Cost-Aware Optimal Pairwise Pure Exploration

Di Wu
University of Virginia

Chengshuai Shi
University of Virginia

Ruida Zhou
University of California,
Los Angeles

Cong Shen
University of Virginia

Abstract

Pure exploration is one of the fundamental problems in multi-armed bandits (MAB). However, existing works mostly focus on specific pure exploration tasks, without a holistic view of the general pure exploration problem. This work fills this gap by introducing a versatile framework to study pure exploration, with a focus on identifying the pairwise relationships between targeted arm pairs. Moreover, unlike existing works that only optimize the stopping time (i.e., sample complexity), this work considers that arms are associated with potentially different costs and targets at optimizing the cumulative cost that occurred during learning. Under the general framework of pairwise pure exploration with arm-specific costs, a performance lower bound is derived. Then, a novel algorithm, termed CAET (**C**ost-**A**ware Pairwise **E**xploration **T**ask), is proposed. CAET builds on the track-and-stop principle with a novel design to handle the arm-specific costs, which can potentially be zero and thus represent a very challenging case. Theoretical analyses prove that the performance of CAET approaches the lower bound asymptotically. Special cases are further discussed, including an extension to regret minimization, which is another major focus of MAB. The effectiveness and efficiency of CAET are also verified through experimental results under various settings.

1 INTRODUCTION

The study of multi-armed bandits (MAB), which captures the essence of sequential decision-making processes, has a long and rich history (Thompson, 1933; Lattimore and Szepesvári, 2020). Despite being a simple model, MAB has found wide-ranged applications including online recommendations (Li et al., 2010), wireless communications (Gai et al., 2010), resource allocation (Liu et al., 2021), clinical trails (Aziz et al., 2021), etc. The targets of learning in MAB can be generally categorized as regret minimization (Auer et al., 2002) and pure exploration (Audibert and Bubeck, 2010).

The most classical pure exploration setting is best arm identification (BAI) (Garivier and Kaufmann, 2016; Kaufmann et al., 2016), whose extensions have also been broadly studied, e.g., top arms identification (Kalyanakrishnan et al., 2012; Zhou and Tian, 2022). However, taking a closer look at these studies on pure exploration, we recognize that there are two major limitations. First, they are all confined to one specific kind of pure exploration task, while a general framework is still lacking. As a result, whenever a different task is of interest (e.g., identifying the ranking of the arms according to their expected rewards), a new algorithm needs to be developed. Second, they commonly treat all arms equally by setting the target as minimizing the total times of arm pulling (i.e., sample complexity). However, in real-world applications, the costs of sampling different arms may differ. For example, in clinical trials (which is further discussed in Sec. 3.4), the prices of varying candidate vaccines or treatments are mostly different from each other.

This work is motivated by these limitations and targets to provide a more comprehensive study of the pure exploration problem. The detailed contributions are summarized as follows:

- A general framework is established to study pure exploration tasks, with a focus on identifying the pairwise relationship between arms. Furthermore, arm-specific costs are introduced so that the ulti-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

mate target is to minimize the cumulative costs during learning. This framework is versatile, subsuming many representative pure exploration tasks, e.g., best arm identification and ranking identification.

- Through the lens of this framework, unified studies on pure exploration are performed. Importantly, a generic lower bound is derived to characterize the fundamental learning limit, capturing the optimal proportions of costs to be assigned to each arm.
- Under the pairwise pure exploration framework, a novel δ-PAC algorithm, termed CAET (**C**ost-**A**ware Pairwise **E**xploration **T**ask), is proposed. It builds on the philosophy of track-and-stop from the study of BAI (Garivier and Kaufmann, 2016) with specifically crafted designs. In particular, a novel *forced exploration* method is introduced so that the sampling proportion between arms with zero and non-zero costs can be carefully balanced. We note that none of the previous works have tackled a pure exploration problem with zero-cost arms, to the best of our knowledge.
- Theoretical analyses demonstrate that CAET is capable of approaching the lower bound in the asymptotic regime, illustrating its optimality. In particular, to obtain the desired performance guarantee, a strong result on the convergence of the empirical sample proportion to the optimal proportion is established.
- We further extend the discussion to the objective of regret minimization, which under the explore-then-commit scheme can be viewed as a BAI problem with the arms' sub-optimality gaps as costs. We demonstrate that in this case, CAET can still achieve asymptotically optimal performance.

2 RELATED WORKS

In the long history of MAB research, different kinds of pure exploration tasks have been studied, where the problem of **best arm identification (BAI)** is arguably the most basic one. As a result, most of the existing studies focus on BAI, e.g., Even-Dar et al. (2006); Gabillon et al. (2012); Kaufmann et al. (2016); Garivier and Kaufmann (2016); Shen (2019). A few extensions have also been investigated. For example, Kalyanakrishnan et al. (2012); Kalyanakrishnan and Stone (2010); Chen et al. (2017b); Zhou and Tian (2022) have studied the identification of the *top few arms*, instead of only the best one. Also, Gabillon et al. (2011) has initiated the investigation on identifying the optimal arm in each pre-specified arm group (which is often referred to as the multi-bandit BAI). However, as mentioned in Section 1, all of these works are confined to one specifically formulated pure exploration problem, while our work provides a general framework that can subsume the previous settings. In addition, most existing works, including the above-mentioned ones, target at optimizing the sample complexity, where all arm pulls are treated equally. This work, however, considers that the arms are associated with potentially different costs with a target of optimizing the cumulative cost. To the best of our knowledge, this is the first time that a general pure exploration framework with arm-specific cost has been established.

In the following, we discuss a few key related works. Chen et al. (2017a) studied the so-called "general sampling problem" in the setting of pure exploration, and provided a lower and upper bound. However, it is still focusing on the uniform cost and many applications with arm-specific cost will be limited, in the meantime, our work generalizes the bandit's framework to pure exploration with the arm-specific cost and provides the optimal lower bound and a near-optimal algorithm. It is noted that the work of Kanarios et al. (2024) touches upon **arm-specific costs** in BAI. In contrast, our work not only studies a more general framework but also considers more general non-negative costs. Furthermore, the investigation of Kanarios et al. (2024) is restricted to strictly positive costs, which limits its applicability, e.g., it cannot handle regret minimization as our extension in Section 7. We particularly note that the *zero-cost arms* introduce significant challenges for both design and analysis, which will be illustrated in later discussions. Also, the recent work Qin et al. (2025) involves costs in the study of BAI, while focusing on finding the arm with the highest reward-to-cost ratio. Moreover, another recent work of Zhang and Ying (2024) studies how to leverage the explore-then-commit scheme to perform regret minimization, which is similar to the extension in Section 7. The differences between Zhang and Ying (2024) and this work are further discussed in Section 7.

Here we also discuss the setting of **dueling bandits** (Du et al., 2022; Yue et al., 2012; Dudík et al., 2015), highlighting its differences from the pairwise pure exploration task in this work. In particular, for **dueling bandits**, a pairwise *feedback* is obtained by selecting two arms at each time step. The pairwise pure exploration task studied in our work is still under the canonical bandit setting, where an arm is pulled at each time step. The reason we name it a pairwise exploration task is that many pure exploration tasks can be exactly represented via pairwise comparisons. For example, in the BAI problem, to guarantee that arm a^* is the best arm, it suffices to show that arm a^* has a larger reward in all $K-1$ pairwise comparisons between arms (a^*, a) for each $a \neq a^*$. To this end, to tackle the pure exploration tasks, it is sufficient to conduct these elementary pairwise comparisons. The class of tasks studied in this paper – *pairwise pure exploration task* – is rigorously defined in Definition 1.

3 PROBLEM FORMULATION

3.1 Bandits with Arm-Dependent Costs

In this work, we consider a multi-armed bandits (MAB) model with a set of K arms, denoted as $\mathcal{A} = \{1, 2, \dots, K\}$. As in the canonical MAB studies, each arm $a \in \mathcal{A}$ is associated with a reward distribution π_a whose expectation is denoted as μ_a. Moreover, each arm $a \in \mathcal{A}$ is associated with a cost distribution ν_a, whose expectation is denoted as c_a. To reflect the nature of costs, we assume that $c_a \geq 0$ for all $a \in \mathcal{A}$. The introduction of the cost distribution allows us to define a broad class of cost-aware pairwise pure exploration tasks later. To ease the exposition, we denote $\boldsymbol{\mu} := (\mu_1, \mu_2, \dots, \mu_K)$ and $\boldsymbol{c} := (c_1, c_2, \dots, c_K)$.

In this model, at each time t, an agent plays an arm $A_t \in \mathcal{A}$. Then, she simultaneously receives a reward X_t and cost value C_t independently sampled from the probability distribution π_{A_t} and ν_{A_t} respectively.

In the following discussion, we focus on the standard scenario (Garivier and Kaufmann, 2016) with the reward distributions belonging to a canonical exponential family, defined as

$$\mathcal{P} = \left\{ (\pi_\theta)_{\theta \in \Theta} : \frac{d\pi_\theta}{d\xi} = \exp(\theta x - b(\theta)) \right\},$$

where $\Theta \in \mathbb{R}$, ξ is some reference measure on $\mathbb{R}$, and $b : \Theta \to \mathbb{R}$ is a convex, twice differentiable function. A distribution $\pi_\theta \in \mathcal{P}$ can be parameterized by its expectation $\dot{b}(\theta)$ and for every $\mu \in \dot{b}(\Theta)$, we denote by π^μ the unique distribution in $\mathcal{P}$ with expectation μ. Then, the bandit model with reward distributions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ can be represented by its means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$.

3.2 The Pairwise Pure Exploration Problem

Denote by S_K the permutation group of $\mathcal{A}$, a pure exploration task $(\mathcal{G}, \varphi)$ can be defined through a partition $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_M\}$ of S_K (i.e., $\mathcal{G}_i \cap \ \mathcal{G}_j = \emptyset$ and $\bigcup_{m=1}^M \mathcal{G}_m = S_K$) and a mapping $\varphi : \mathbb{R}^K \to [M]$ that outputs an index within the partition $\varphi(\boldsymbol{\mu}) = m \in [M]$ for any expected reward vector $\boldsymbol{\mu} \in \mathbb{R}^K$.

We are interested in a sufficiently expressive and representative subclass of all the pure exploration tasks, referred to as the **pairwise exploration tasks**. To ease the notation, denote by $\sigma^{\boldsymbol{\mu}}$ the index of descending order of $\boldsymbol{\mu}$, i.e., $\boldsymbol{\mu}_{\sigma^{\boldsymbol{\mu}}(1)} > \boldsymbol{\mu}_{\sigma^{\boldsymbol{\mu}}(2)} > \cdots > \boldsymbol{\mu}_{\sigma^{\boldsymbol{\mu}}(K)}$.[1] Then, $(\sigma^{\boldsymbol{\mu}})^{-1}(j) = k$ indicates that arm-j of $\boldsymbol{\mu}$ is the k-th largest arm, and $(\sigma^{\boldsymbol{\mu}})^{-1}(j) < (\sigma^{\boldsymbol{\mu}})^{-1}(i)$ indicates the reward of arm-j is larger than that of arm-i. Let $\mathcal{B}_{ij} := \{\sigma \in S_K : \sigma^{-1}(i) < \sigma^{-1}(j)\}$. The pairwise exploration task can be defined as follows:

Definition 1 (Pairwise Exploration Task) *A task $(\mathcal{G}, \varphi)$ is a pairwise exploration task if for any $\mathcal{G}_m \in \mathcal{G}$, there exists a subset $\mathcal{I}_m \subseteq \{(i,j) : 1 \leq i \neq j \leq K\}$ such that $\mathcal{G}_m = \cap_{(i,j)\in\mathcal{I}_m} \mathcal{B}_{ij}$ and $\varphi(\boldsymbol{\mu}) = m$ if $\sigma^{\boldsymbol{\mu}} \in \mathcal{G}_m$.*

Intuitively, we can interpret the pairwise exploration task as the task that can be solved via confirming a set of binary comparisons among $\boldsymbol{\mu}$. The intersection $\cap_{(i,j)\in\mathcal{I}} \mathcal{B}_{ij}$ indicates checking the condition of $\wedge_{(i,j)\in\mathcal{I}} (\sigma^{\boldsymbol{\mu}}(i) < \sigma^{\boldsymbol{\mu}}(j)) = \wedge_{(i,j)\in\mathcal{I}} \{\text{arm-}i \text{ is better than arm-}j \text{ in } \boldsymbol{\mu}\}$.

The class of pairwise exploration tasks are sufficiently large that contains many tasks of interest as special cases. We give some examples below.

Example 1 (Ranking Identification) *Considering the pure exploration problem $(\mathcal{G}, \varphi)$ of identifying the rank of all the arms with respect to their expected rewards. Thus $M = K!$ and each $\mathcal{G}_m \in \mathcal{G}$ is a singleton containing some σ. It is a pairwise exploration task since $\{\sigma\} = \cap_{i=1}^{K-1} \mathcal{B}_{\sigma^{-1}(i)\sigma^{-1}(i+1)}$.*

It can be observed that ranking identification is a particularly challenging task because $M = K!$, even though the minimum number of binary comparisons needed to identify a ranking is only $K - 1$.

Example 2 (BAI) *In the pure exploration problem $(\mathcal{G}, \varphi)$ of identifying the best arm, $M = K$ and $\mathcal{G}_k = \cap_{j\neq k} \mathcal{B}_{kj}$, which is a pairwise exploration task.*

Example 3 (Best-m-arms Identification) *Considering the pairwise pure exploration problem $(\mathcal{G}, \varphi)$ of identifying the set of m arms with the highest rewards. We can take $M = \binom{K}{m}$ and for each $\boldsymbol{i} = (i_1, \dots, i_m)$ and $\mathcal{G}_{\boldsymbol{i}} = \cap_{k\notin\{i_1,\dots,i_m\}} \cap_{t=1}^m \mathcal{B}_{i_t k}$.*

Since for pairwise exploration tasks $(\mathcal{G}, \varphi)$, φ is defined explicitly by $\mathcal{G}$, we omit φ and use $\mathcal{G}$ to represent the exploration tasks for simplicity.

3.3 Cost-Aware Pure Exploration with Fixed Confidence

Denoting $\mathcal{F}_t = \sigma(X_1, C_1, \dots, X_t, C_t)$ as the σ-field generated by the observations (i.e., both rewards and

[1] We mainly focus on the $\boldsymbol{\mu}$ without ties, and discuss the case with ties for different tasks individually.

costs) up to time t, a strategy should generally have the following three components:

- a sampling rule $(A_t)_t$, where A_t is $\mathcal{F}_t$ measurable;
- a stopping rule τ, where τ is a stopping time with respected to $\mathcal{F}_t$; and
- an $\mathcal{F}_\tau$-measurable decision rule $\hat{m}_\tau$.

This work focuses on the so-called *fixed-confidence* setting and targets at minimizing the cumulative costs. Specifically, given a confidence parameter $\delta \in [0, 1]$, the designed strategy should guarantee that the decision rule outputs the correct index m with probability at least $1 - \delta$, i.e.,

$$\mathbb{P}(\hat{m}_\tau \neq m) \leq \delta,$$

while minimizing the expected incurring cost defined as

$$\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau)] = \sum_{a \in \mathcal{A}} c_a \mathbb{E}[N_a(\tau)], \tag{1}$$

where $N_a(\tau) = \sum_{t \leq \tau} \mathbb{1}\{A_t = a\}$. As the stopping rule τ in general would be related to the required confidence δ, we adopt the notation τ_δ to specify this dependency.

It can be realized that the canonical study of pure exploration problems targets at ensuring the identification correctness while minimizing the sample complexity τ, which can be interpreted as a specific case of an all-one cost vector $\boldsymbol{c} = (1, \ldots, 1)$ in Eqn. (1). The consideration of the general pure exploration tasks and the flexible cost vector $\boldsymbol{c}$ largely broaden the scope of the previous pure exploration studies. In particular, there may exist $c_i = 0$ in the cost vector $\boldsymbol{c}$, which significantly increases the technical difficulty in the strategy design and analysis, as will be evident later.

Preliminaries and Notations For two probability distributions p and q, we denote their Kullback-Leibler (KL) divergence as $\mathrm{KL}(p, q)$. As has been stated in Cappé et al. (2013), the KL divergence from two distributions π_θ and $\pi_{\theta'}$ in the exponential family, with expectations μ and μ', induces a divergence function d on $\dot{b}(\Theta)$ defined as

$$d(\mu, \mu') = \mathrm{KL}(\pi_\theta, \pi_{\theta'}) = b(\theta') - b(\theta) - \dot{b}(\theta)(\theta' - \theta).$$

Specifically, when two reward distributions are Gaussian with expectations x, y and variance σ^2, it holds that $d(x, y) = (x - y)^2/(2\sigma^2)$; when two reward distributions are Bernoulli with expectations x, y, it holds that $d(x, y) = \mathrm{kl}(x, y) = x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$.

We denote $P(\boldsymbol{c}) := \{a \in \mathcal{A} : c_a > 0\}$ as the set of arms associated with non-zero expected costs and $N(\boldsymbol{c}) := \{a \in \mathcal{A} : c_a = 0\}$ as the set of arms associated with zero expected costs. Also, for any set $\mathcal{I}$ defined in Definition 1, we define its support set $\mathrm{supp}(\mathcal{I}) = \{i : i \in \mathcal{A}, \exists j \in \mathcal{A}/\{i\}, \text{ s.t. } (i, j) \in \mathcal{I}\}$. For exploration task $(\mathcal{G}, \varphi), \mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}$ and we denote $\mathcal{I}_m$ be the index set of intersection: $\mathcal{G}_m = \cap_{(i,j) \in \mathcal{I}_m} \mathcal{B}_{ij}$. For bandit model $\boldsymbol{\mu}$, with $\varphi(\boldsymbol{\mu}) = m$, we further define $P_\mathcal{G}(\boldsymbol{c}, \boldsymbol{\mu}) = P(\boldsymbol{c}) \cap \mathrm{supp}(\mathcal{I}_m)$, $N_\mathcal{G}(\boldsymbol{c}, \boldsymbol{\mu}) = N(\boldsymbol{c}) \cap \mathrm{supp}(\mathcal{I}_m)$, and $\mathcal{I}(\boldsymbol{\mu}) = \mathcal{I}_m = \mathcal{I}_{\varphi(\boldsymbol{\mu})}$. Furthermore, a notation table is provided in Appendix A to facilitate reading.

3.4 Motivation Applications

In the following, two motivation applications are discussed to highlight the practical relevance of the cost-aware pure exploration problem formulated above.

First, in clinical trials (e.g., identifying the most effective vaccine or treatment among multiple candidates), the costs of different vaccines or treatments can vary significantly. Therefore, adopting a cost-aware design, as proposed in this work, is advantageous, as it can lead to substantial budget savings compared to previous approaches that assume uniform costs. In addition, self-healing (using placebo or no active intervention) often serves as a crucial baseline for evaluating the efficacy of medical treatments. Such baselines inherently have no additional cost and can be modeled effectively as zero-cost arms as considered in this work.

Similarly, in wireless communications, different channels allow for varying communication rates, while also requiring different transmit power levels to communicate. In the task of finding the best channel with the highest communication rate, this work can further minimize the overall power (i.e., cumulative costs) consumed during the process, instead of simply measuring the overall channel usage as in previous works.

4 LOWER BOUND

In this section, we establish the fundamental limits of the general pure exploration problems. Let $(\mathcal{G}, \varphi)$ be the pairwise exploration task that is concerned. We first define $\mathcal{S}$ as the set of bandit models whose expected rewards $\boldsymbol{\mu}$ satisfying that $\varphi(\boldsymbol{\mu})$ belongs to exactly one partition $\mathcal{G}_m$. We say a strategy is δ-PAC if, for every $\boldsymbol{\mu} \in \mathcal{S}$, it satisfies that $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty) = 1$ and $\mathbb{P}_{\boldsymbol{\mu}}(\hat{m}_{\tau_\delta} \neq m) \leq \delta$. Furthermore, we introduce the notion of an *alternative set* as

$$\begin{aligned}\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in S : \varphi(\boldsymbol{\lambda}) \neq \varphi(\boldsymbol{\mu}), \\ \lambda_i = \mu_i, \forall i \notin P_\mathcal{G}(\boldsymbol{c}, \boldsymbol{\mu})\},\end{aligned} \tag{2}$$

and denote the following distribution set

$$\Sigma_{P_\mathcal{G}(\boldsymbol{c}, \boldsymbol{\mu})} := \{\boldsymbol{\omega} \in \Delta_K : \omega_i = 0, \omega_j > 0,$$

$$\forall i \notin P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}), \forall j \in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})\}, \quad (3)$$

where Δ_K denotes the probability simplex of $K-1$ dimensions. Then, the following lower bound can be established.

Theorem 1 *Let $\delta \in (0,1)$ and give a pairwise exploration task $(\mathcal{G},\varphi)$. For any δ-PAC strategy and any bandit model $\boldsymbol{\mu} \in S$, the following holds:*

$$\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)] \geq T^*(\boldsymbol{c},\boldsymbol{\mu})\mathrm{kl}(\delta,1-\delta),$$

where

$$T^*(\boldsymbol{c},\boldsymbol{\mu})^{-1} = \sup_{\omega\in\Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})} \left\{ \sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \omega_a \frac{d(\mu_a,\lambda_a)}{c_a} \right\}. \quad (4)$$

As $\mathrm{kl}(\delta,1-\delta) \sim \log(1/\delta)$ when δ goes to zero, this theorem yields the following asymptotic lower bound:

$$\liminf_{\delta\to 0} \frac{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]}{\log(1/\delta)} \geq T^*(\boldsymbol{c},\boldsymbol{\mu}).$$

A non-asymptotic version can be obtained from the inequality $\mathrm{kl}(\delta,1-\delta) \geq \log(1/(2.4\delta))$ as $\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)] \geq T^*(\boldsymbol{c},\boldsymbol{\mu})\log(1/(2.4\delta))$.

Proof sketch: To establish the lower bound, we first consider a transportation lemma of Kaufmann et al. (2016) which relates to the expected number of draws in two bandit models being different under the given exploration task:

$$\forall\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu}), \sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} d(\mu_a,\lambda_a)\mathbb{E}[N_a(\tau_\delta)] \geq \mathrm{kl}(\delta,1-\delta),$$

Considering $\mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)] := \sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} c_a\mathbb{E}[N_a(\tau_\delta)]$, we have

$$\begin{aligned}\mathrm{kl}(\delta,1-\delta) &\leq \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})} \mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]\cdot \left(\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \frac{c_a\mathbb{E}[N_a]}{\mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]} \frac{d(\mu_a,\lambda_a)}{c_a} \right) \\ &\leq \mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]\cdot \sup_{\omega\in\Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})} \left(\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \omega_a \frac{d(\mu_a,\lambda_a)}{c_a} \right).\end{aligned}$$

Finally, we have the desired result $\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)] \geq \mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)] \geq T^*(\boldsymbol{c},\boldsymbol{\mu})\mathrm{kl}(\delta,1-\delta)$. ■

It can be observed that the supremum in $T^*(\boldsymbol{c},\boldsymbol{\mu})$ is indeed a maximum because, by Lemma 3, $\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})}\{\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}\omega_a d(\mu_a,\lambda_a)/c_a\}$ is a continuous function, and any continuous function defined in a closed set can reach its maximum value. Thus, we define

$$\boldsymbol{\omega}^*(\boldsymbol{c},\boldsymbol{\mu}) := \operatorname*{arg\,max}_{\boldsymbol{\omega}\in\Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})} \left\{ \sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \omega_a \frac{d(\mu_a,\lambda_a)}{c_a} \right\},$$

which is directly related to the optimal sampling proportion and ω_a represents the ratio of arm a among all the arms. However, the vector $\boldsymbol{\omega}^*$ is not the desired sample distribution due to the scaling of the cost vector. Thus, the following transform function $G_{\boldsymbol{c}} : \Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \to \Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}$ is defined to get the actual sampling distribution from $\boldsymbol{\omega}^*$: for cost vector $\boldsymbol{c}$,

$$G_{\boldsymbol{c}}([a_1,\ldots,a_K]) = [b_1,\ldots,b_K], \text{where}$$
$$b_i = \begin{cases} \frac{a_i/c_i}{\sum_{s\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} a_s/c_s} & \text{if } i\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}) \\ 0 & \text{if } i\notin P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}). \end{cases}$$

The complete proof of Theorem 1, which is given in Appendix C.1, shows that $\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu}) := G_{\boldsymbol{c}}(\boldsymbol{\omega}^*(\boldsymbol{c},\boldsymbol{\mu}))$ is the proportion of pulling arms in $P(\boldsymbol{c}) \cap \mathrm{supp}(\mathcal{I}_m)$ that matches this lower bound. Thus, the pulling proportion of the arm in $P(\boldsymbol{c})$ in the optimal algorithm should be with respect to this pulling proportion. Our sampling rule in Section 5 indeed follows this intuition. Additional results on more explicit forms of $T^*(\boldsymbol{c},\boldsymbol{\mu})$ can be found in Appendix C.2.

5 CAET ALGORITHM

We now describe a novel strategy to tackle the cost-aware pairwise exploration problem $\mathcal{I}$ called CAET (**C**ost-**A**ware Pairwise Pure **E**xploration **T**ask). Without loss of generality, we consider the scenarios with $\mathrm{supp}(\mathcal{I}) = \mathcal{A}$, as otherwise one can exclude the arms not considered in $\mathcal{I}$ from $\mathcal{A}$ to form a new arm set. We will present a sampling rule in Section 5.1, a stopping rule in Section 5.2, and a decision rule in Section 5.3, whose combination leads to the proposed CAET algorithm that is compactly presented in Algorithm 1. In the following discussion, the set $N(\boldsymbol{c})$ will be assigned in advance, and we generally denote the current sample means of expected rewards $\boldsymbol{\mu}$ at time t as $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t),\ldots,\hat{\mu}_K(t))$ where $\hat{\mu}_a(t) = N_a(t)^{-1}\sum_{s\leq t} X_s\mathbf{1}_{\{A_s=a\}}$ and the current sample means of expected costs $\boldsymbol{c}$ at time t as $\hat{\boldsymbol{c}}(t) = (\hat{c}_1(t),\ldots,\hat{c}_K(t))$ where $\hat{c}_a(t) = N_a(t)^{-1}\sum_{s\leq t} C_s\mathbf{1}_{\{A_s=a\}}$ for arm $a\in P(\boldsymbol{c})$.

5.1 Sampling Rule

We begin the description of the sampling rule with the simple scenario where the expected costs are all positive, i.e., $c_a > 0$ for all $a \in \mathcal{A}$, or equivalently, $|N(\boldsymbol{c})| = 0$. As the optimal sampling proportion for all arms in this case is $\boldsymbol{u}^*(\boldsymbol{c}, \boldsymbol{\mu}) = G_{\boldsymbol{c}}(\boldsymbol{\omega}^*(\boldsymbol{c}, \boldsymbol{\mu}))$, one natural idea is to sample according to the plug-in estimate $\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$. Especially, the C-tracking strategy proposed in Garivier and Kaufmann (2016) can be adopted.

With the sampling intuition explained under non-zero expected costs, we move to the challenging case with some arms having zero expected costs, i.e., $|N(\boldsymbol{c})| \neq 0$. It can be observed that the introduction of zero costs drastically complicates the problem, as $\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$ (cf. step 4 in Algorithm 1) only has non-zero elements for arms in $P(\boldsymbol{c})$ and thus does not describe how arms in $N(\boldsymbol{c})$ should be sampled. To overcome this issue and particularly guarantee a finite stopping time, we introduce a proportion $\alpha \in (0, 1)$ to perform uniform sampling over arms with zero costs. The arms with non-zero costs share the other $1 - \alpha$ pulling proportion and are still sampled by tracking the estimate of the optimal proportion as described before. In particular, we design the following overall distribution to be tracked (cf. step 5 in Algorithm 1):[2]

$$\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) = \underbrace{\mathbf{1}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))}_{\text{arms in } N_{\mathcal{G}}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))} + \underbrace{(1-\alpha)\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))}_{\text{arms in } P_{\mathcal{G}}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))},$$

where $\boldsymbol{c}_\delta(t) = D_\delta(\hat{\boldsymbol{c}}(t))$, $D_\delta([a_1, \dots, a_K]) = [b_1, \dots, b_K]$ with $b_i = a_i$ if $a_i > \gamma_0 \log^{-r'}(1/\delta)$ and $b_i = 0$ otherwise, where γ_0 is an arbitrary positive constant and r' is a constant satisfying $0 < r' < 1/8$. Also, $\mathbf{1}_\alpha(\boldsymbol{c}', \boldsymbol{\mu}') := (i_1, \dots, i_K)$ with $i_t = \alpha / |N_{\mathcal{G}}(\boldsymbol{c}', \boldsymbol{\mu}')|$ if $t \in N_{\mathcal{G}}(\boldsymbol{c}', \boldsymbol{\mu}')$ and $i_t = 0$ otherwise. It can be observed that the parameter α balances the sampling proportion between the arms with zero and non-zero costs. Intuitively, to guarantee a small cumulative cost when $\delta \to 0$, those arms with zero costs should be pulled sufficiently so that their estimations are precise enough to guide the estimation of $\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$ for the other arms. Thus, intuitively, the choice of α should be adaptive in δ such that $\lim_{\delta \to 0} \alpha = 1$. In particular, we take $\alpha = 1 - \log^{-r}(1/\delta)$ in the proposed CAET design for any chosen $0 < r < 1/2$. More detailed explanations are provided in Section 6.1.

Algorithm 1: CAET (**C**ost-**A**ware Pairwise Pure **E**xploration **T**ask)

```
1  Input: $K$, confidence $\delta$, $\boldsymbol{\mu}, \boldsymbol{\nu}$;
2  while $t \in \mathbb{N}$, $\exists (a, b) \in \mathcal{I}$ s.t., $Z_{a,b}(t) \leq \beta(t, \delta)$ do
3      Calculate $\hat{\boldsymbol{c}}(s), \hat{\boldsymbol{\mu}}(s), s \leq t$;
4      Calculate
         $\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) = G_{\boldsymbol{c}_\delta(t)}(\boldsymbol{\omega}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)))$;
5      Calculate $\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) =$
         $\mathbf{1}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) + (1-\alpha)\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$;
6      Choose action $A_{t+1} \in$
         $\arg\max_{1 \leq a \leq K} \sum_{s=0}^{t} u^{\epsilon_s}_{\alpha,a}(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s)) - N_a(t)$,;

7      Observe $X_{t+1}$ and $C_{t+1}$ and update data;
8      $t = t + 1$;
9  end
10 Output: Output the $\mathcal{G}_m$ determined by the
     stopping rule Equation (6)
```

Finally, the sampling rule (cf. step 6 in Algorithm 1) is designed as

$$A_{t+1} \in \operatorname*{arg\,max}_{1 \leq a \leq K} \sum_{s=0}^{t} u^{\epsilon_s}_{\alpha,a}(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s)) - N_a(t), \qquad (5)$$

where $\boldsymbol{u}^\epsilon_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) = (u^\epsilon_{\alpha,1}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)), \dots, u^\epsilon_{\alpha,K}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)))$ is the L^∞ projection of $\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$ to $\Sigma^\epsilon_K = \{(\omega_1, \dots, \omega_K) \in [\epsilon, 1]^K : \sum_{i=1}^K \omega_i = 1\}$. This sampling rule guarantees that besides some forced explorations for information collection introduced by the projection, the arms are pulled roughly according to $\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$. For the choice of ϵ_t, we take $\epsilon_t = (K^2 + t)^{-1/2}/2$, which leads to the desired performance. One important property of this sampling rule is that the empirical sample proportion is guaranteed to converge to the optimal proportion, which will be discussed in Section 6.1.

5.2 Stopping Rule

For any pair of arms $a, b \in \mathcal{A}$, we consider the *generalized likelihood ratio* as follows:

$$Z_{a,b}(t) := \log \frac{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{X}^a_{N_a(t)}) p_{\mu'_b}(\underline{X}^b_{N_b(t)})}{\max_{\mu'_a \leq \mu'_b} p_{\mu'_a}(\underline{X}^a_{N_a(t)}) p_{\mu'_b}(\underline{X}^b_{N_b(t)})},$$

where $\underline{X}^a_{N_a(t)} = (X_s : A_s = a, s \leq t)$ contains the observations of arm a available at time t, and $p_\mu(Z_1, \dots, Z_n)$ is the likelihood of n i.i.d. observations from ω^μ as

$$p_\mu(Z_1, \dots, Z_n) = \prod_{k=1}^{n} \exp(\dot{b}^{-1}(\mu) Z_k - b(\dot{b}^{-1}(\mu))).$$

For the exponential bandit models, if $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$, we have

$$Z_{a,b}(t) = N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{a,b}(t)) + N_b(t) d(\hat{\mu}_b(t), \hat{\mu}_{a,b}(t)),$$

[2] When there is no cost-zero arm, i.e. $N_{\mathcal{G}}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) = \emptyset$, the sample distribution can be directly set as $\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$.

where

$$\hat{\mu}_{a,b}(t) := \frac{N_a(t)}{N_a(t)+N_b(t)}\hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t)+N_b(t)}\hat{\mu}_b(t)$$

is the empirical mean of arms a, b. Note that $Z_{a,b}(t)$ is non-negative if and only if $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$. Thus, in order to identify to which partition $\mathcal{G}_m$ $\boldsymbol{\mu}$ belongs, we use $\mathcal{I}_m$ to represent the index set $\mathcal{I}$ of $\mathcal{G}_m = \cap_{(i,j)\in\mathcal{I}_m}\mathcal{B}_{ij}$ and check the value of $Z_{a,b}(t)$. The stopping rule can be specified as

$$\begin{aligned}&\tau_\delta = \\ &\inf\{t \in N : \exists \mathcal{G}_m, \forall (a,b) \in \mathcal{I}_m, Z_{a,b}(t) > \beta(t,\delta)\},\end{aligned} \tag{6}$$

where the threshold $\beta(t,\delta)$ needs a careful design to ensure the δ-PAC property. Several provably efficient choices of $\beta(t,\delta)$ are provided in Section 6.2 with the corresponding analyses.

5.3 Decision Rule

When the algorithm stops, it means there exists $\mathcal{G}_m$ such that $\forall (a,b) \in \mathcal{I}_m, Z_{a,b}(t) > \beta(t,\delta)$ at a certain time τ_δ. Our decision rule $\hat{m}_\tau$ will output this $\mathcal{G}_m$.

6 UPPER BOUND

In this section, we first analyze the convergence of the empirical sample proportion to the optimal one. Then, a few provably efficient choices of the threshold in the stopping rule are provided. Finally, the expected asymptotic convergence of CAET is presented, which approaches the lower bound in Theorem 1, demonstrating the asymptotic optimality of CAET.

6.1 Convergence to the Optimal Proportion

First, we discuss the convergence of the empirical sampling proportion to the optimal proportion, where the existence of zero-cost arms introduces analytical challenges. Especially, when $N_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}) = \emptyset$, the optimal proportion $\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu})$ does not change with the choice of δ. Since $\lim_{\delta\to 0}\tau_\delta = \infty$, Proposition 3 is sufficient to obtain the convergence guarantee.

However, when there are zero-cost arms in $\mathrm{supp}(\mathcal{I}(\boldsymbol{\mu}))$, $N_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}) \neq \emptyset$, and the optimal proportion $\boldsymbol{u}_\alpha(\boldsymbol{c},\boldsymbol{\mu}) = \mathbf{1}_\alpha(\boldsymbol{c}) + (1-\alpha)\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu})$ change with δ due to the δ-dependency of the choice of α. As a result, the key challenge is to prove that

$$\mathbb{P}\left(\lim_{\delta\to 0}\frac{N_a(\tau_\delta)}{(1-\alpha)\tau_\delta} = u_a^*\right) = 1, \quad a \in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu});$$

$$\mathbb{P}\left(\lim_{\delta\to 0}\frac{N_a(\tau_\delta)}{\tau_\delta} = \frac{1}{|N(\boldsymbol{c})|}\right) = 1, \quad a \in N_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})$$

Moreover, for arm $a \in P(\boldsymbol{c})$, the optimal proportion $u_a = (1-\alpha)u_a^*$ will go to zero with $\alpha \to 1$ when $\delta \to 0$, which means the time of u_a under the influence of the ϵ^t-L^∞ projection keeps increasing. Thus, the main difficulty in proving the desired convergence is to handle the balance between the stopping time τ_δ and the amount of time that L^∞ has an impact on u_a. To tackle the problem, we derive the potential proprieties of a series of random processes and use some adaptive sampling technology to control the whole process, which is novel and of its own merit. The detailed proof is provided in Proposition 7 of Appendix D.

6.2 Threshold in the Stopping Rule

Parameter $\beta(t,\delta)$ can be viewed as an exploration rate, which requires a careful design to achieve the δ-PAC property. In the following, two provably efficient methods are provided.

Proposition 1 (Informational Threshold) *Let $\boldsymbol{\mu}$ be a Bernoulli bandit model. Let $\delta \in (0,1)$. For any sampling strategy, using the stopping rule given in Equation (6) on Bernoulli bandits with threshold $\beta(t,\delta) = \log(2tK(K-1)/\delta)$ ensures that for all $\boldsymbol{\mu} \in S$, $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty, \varphi(\hat{\boldsymbol{\mu}}(\tau_\delta)) \neq \varphi(\boldsymbol{\mu})) \leq \delta$.*

For the deviational threshold, to guarantee the δ-PAC property in any exponential bandit models, by making use of a deviation result from Magureanu et al. (2014), we have the following proposition.

Proposition 2 (Deviational Threshold) *Let $\boldsymbol{\mu}$ be an exponential family of bandit models. Let $\delta \in (0,1)$ and $\theta > 1$. There exists a constant $C = C(\theta, K)$ such that for any sampling strategy, using the stopping rule given in Equation (6) with threshold $\beta(t,\delta) = \log\left(Ct^\theta/\delta\right)$ ensures that for all $\boldsymbol{\mu} \in S$, $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty, \varphi(\hat{\boldsymbol{\mu}}(\tau_\delta)) \neq \varphi(\boldsymbol{\mu})) \leq \delta$.*

The proof of these two propositions can be found in Appendix E.

6.3 Asymptotic Optimality in Expectation

The following theorem establishes the asymptotic optimality of the proposed CAET algorithm.

Theorem 2 (Upper Bound) *Let $\theta \in [1, e/2]$ and $r(t) = O(t^\theta)$. Using the stopping rule in Equation (6) with an exploration rate $\beta(t,\delta) = \log(r(t)/\delta)$ and the sampling rule in Equation (5) with a sample distribution function $\boldsymbol{u}_\alpha$ under $\alpha = 1 - \log^{-r}(1/\delta)$, we have*

$$\limsup_{\delta\to 0}\frac{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]}{\log(1/\delta)} \leq \theta T^*(\boldsymbol{c},\boldsymbol{\mu}).$$

Proof sketch: We provide a brief proof sketch of Theorem 2, while the complete proof is given in Appendix 6. Intuitively, directly calculating the cumulative cost $\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)]$ is hard, since it needs to handle the arm-specific costs. Thus, by extending Garivier and Kaufmann (2016), we can first obtain the following result regarding the sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$:

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \theta T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu}). \tag{7}$$

Then a key result in Lemma 1 is established to transfer $\mathbb{E}_\mu[\tau_\delta]$ to $\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)]$, which reveals the deep relationship between the cumulative cost and the sample complexity.

Lemma 1 *Using the CAET algorithm with the optimal sampling proportion set as $\boldsymbol{u}^*_\alpha = \boldsymbol{u}_\alpha(\boldsymbol{c}, \boldsymbol{\mu})$, when $\delta \to 0$, the complexity $T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu})$ in Equation* (7) *satisfies*

$$\lim_{\delta \to 0} T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu})^{-1} \frac{\mathbb{E}_\mu[\tau_\delta]}{\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)]} = T^*(\boldsymbol{c}, \boldsymbol{\mu})^{-1}.$$

Thus, it can be observed that for Bernoulli bandit $\boldsymbol{\mu}$, the choice $\beta(t, \delta) = \log(2tK(K-1)/\delta)$ in the stopping rule is δ-PAC. Also, with the sampling rule given above, the stopping time τ_δ is almost surely finite. When δ is small enough, its expectation approaches $T^*(\boldsymbol{\mu}) \log(1/\delta)$. Thus, the optimal sample complexity is established in Theorem 1. ■

More generally, for the exponential family bandits, combining Proposition 2 and Theorem 2, it can be obtained that for every $\theta > 1$, there exists an exploration rate such that CAET is δ-PAC and satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)]}{\log(1/\delta)} \leq \theta T^*(\boldsymbol{c}, \boldsymbol{\mu}).$$

Finally, we note that the δ-dependency of the choice of α is related to the convergence rate of sample complexity and the order of stopping time. Especially, a faster convergence rate of α to 1 improves the convergence rate of the algorithm, but at the expense of increasing the stopping time. The proposed CAET algorithm takes $\alpha = 1 - \log^{-r}(1/\delta)$ with $0 < r < 1/2$, achieving stopping time of order $O(\log^{1+r}(1/\delta))$.

7 DISCUSSION ON SPECIAL CASES

The upper bound analyses in Theorem 2 are for the general pairwise pure exploration problem, which can be further refined for various specifically targeted tasks, e.g., best-arm identification and ranking identification. Details on these two specifications are articulated in Appendix G. In this section, we provide additional results on extending the study to the problem of *regret minimization.*

We define the optimal action a^* as the arm with the highest expected reward, i.e., $a^* := \arg\max_{a \in \mathcal{A}} \mu_a$. Also, the sub-optimal gap between arm a and optimal arm a^* is denoted as $\Delta_a := \mu_{a^*} - \mu_a$.

To realize the regret minimization setting, we can track the sub-optimal gap through the difference of respective sample means in each round. In round t, we take $\boldsymbol{\nu}(t) = (0, \nu_{\Delta_2}(t), \nu_{\Delta_3}(t), \ldots, \nu_{\Delta_K}(t))$, where $\nu_{\Delta_i}(t) = \hat{\mu}_1(t) - \hat{\mu}_i(t)$ is the estimation of the sub-optimal gap. This scheme of observing the cost vector slightly differs from the setting described in Section 3 where the costs are always sampled from a certain distribution. However, as shown in the proofs, the proposed CAET algorithm is still functional in this scenario.

Furthermore, the discussions of BAI in Appendix G demonstrate the commitment time of CAET to the optimal arm is of order $O(\log(1/\delta)^{1+r})$. Then, with $\delta = 1/T$ and an explore-then-commit (ETC) scheme (i.e., keep pulling the identified arm after commitment), CAET achieves

$$R_{\boldsymbol{\mu}}(T) \leq \theta T^*(\boldsymbol{\Delta}, \boldsymbol{\mu}) \log(T), \tag{8}$$

$$T^*(\boldsymbol{\Delta}, \boldsymbol{\mu}) = \sum_{\Delta_a > 0} \frac{\Delta_a}{\mathrm{KL}(\mu_a, \mu_1)}, \tag{9}$$

which is further elaborated in Appendix G.3.

Similar ideas of leveraging the ETC scheme to perform regret minimization using BAI designs have been recently studied in Zhang and Ying (2024). For Bernoulli bandits, both CAET and Zhang and Ying (2024) achieve asymptotically optimal regret. For general exponential bandits, CAET is capable of achieving an asymptotically near-optimal performance (with $\theta > 1$) without a pre-determined stopping time, which is required in Zhang and Ying (2024). Note that although the commitment time $O(\log(1/\delta)^{1+r})$ of CAET is slightly larger than the pre-determined stopping time order $O(\log(1/\delta))$ in Zhang and Ying (2024), we can take r close enough to zero to approach their result.

8 NUMERICAL EXPERIMENTS

In this section, numerical experiment results are reported to evaluate the effectiveness and efficiency of the proposed CAET algorithm. The task of ranking identification is considered in the experiments with a focus on three-arm bandits, while the cost vector is set as the sub-optimal gap vector to also achieve the goal of regret minimization. A more explicit theoretical result for this particular task is provided in

Proposition 8 in Appendix G. Fig. 1 reports the cumulative regrets obtained by CAET over two Bernoulli bandit instances, $\boldsymbol{\mu}_1 = (3, 4, 2), r = 0.4, \theta = 1.2$ and $\boldsymbol{\mu}_2 = (1.4, 0.8, 0.3), r = 0.4, \theta = 1.2$ with truncation function $D_\delta(\cdot)$ taking $0.1 \log^{-0.1}(1/\delta)$ as the threshold. The established theoretical lower and upper bounds are also plotted for comparison purposes. It can be observed that when $\delta \to 0$, the performance of CAET is indeed limited by the upper bound and also capable of approaching the lower bound, which corroborates the theoretically asymptotic optimality of CAET.

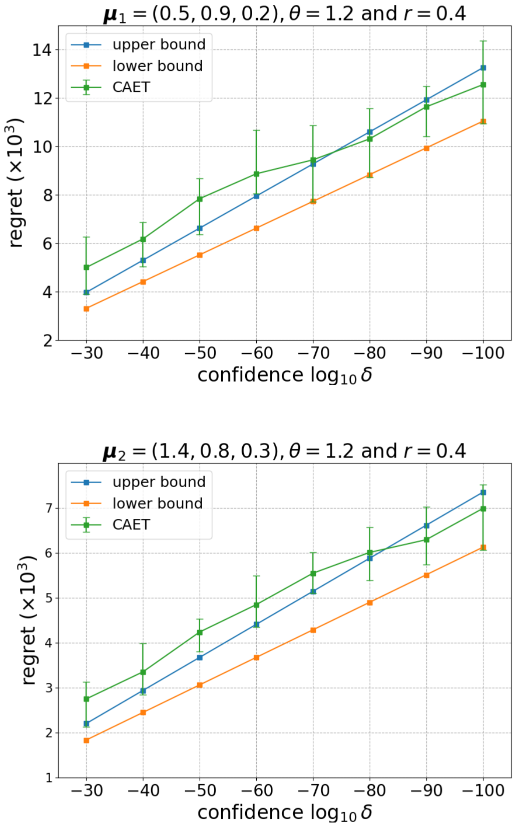

Figure 1: Experimental results

9 CONCLUSIONS

This work introduced a general framework to study pure exploration problems in MAB. It focuses on identifying the pairwise relationships between arm pairs and can flexibly incorporate arm-specific costs to broaden the applicability. A performance lower bound was first established, which characterizes the fundamental limits of learning in the general framework. The novel CAET algorithm was then proposed, incorporating carefully crafted designs to handle arms associated with zero cost. Theoretical analyses demonstrated that CAET is capable of approaching the lower bound asymptotically, highlighting its optimality. An extension to regret minimization via the explore-then-commit scheme was further developed, which provably achieves the optimal performance asymptotically. Experimental results corroborated the effectiveness and efficiency of CAET.

To enhance the applicability of the CAET algorithm, one important future direction is to make it computationally more efficient due to the fact that a complex optimization problem needs to be solved in each round. One feasible solution is to consider a batched version of CAET, whose design and analysis are left for further investigation.

Acknowledgement

The work of D. Wu, C. Shi and C. Shen was partially supported by the U.S. National Science Foundation (NSF) under awards ECCS-2033671, ECCS-2143559, ECCS-2029978, and CNS-2002902. The work of R. Zhou was supported in part by NSF grants 2139304, 2146838 and the Army Research Laboratory grant under Cooperative Agreement W911NF-17-2-0196.

References

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 13–p.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Aziz, M., Kaufmann, E., and Riviere, M.-K. (2021). On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, 22(14):1–38.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullbackleibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3).

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR.

Chen, L., Gupta, A., Li, J., Qiao, M., and Wang, R. (2017a). Nearly optimal sampling algorithms for combinatorial pure exploration.

Chen, L., Li, J., and Qiao, M. (2017b). Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110. PMLR.

Du, Y., Wang, S., and Huang, L. (2022). Dueling bandits: From two-dueling to multi-dueling.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. (2015). Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR.

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(6).

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25.

Gabillon, V., Ghavamzadeh, M., Lazaric, A., and Bubeck, S. (2011). Multi-bandit best arm identification. *Advances in Neural Information Processing Systems*, 24.

Gai, Y., Krishnamachari, B., and Jain, R. (2010). Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN)*, pages 1–9. IEEE.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR.

Jedra, Y. and Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017.

Kalyanakrishnan, S. and Stone, P. (2010). Efficient selection of multiple bandit arms: Theory and practice. In *ICML*, volume 10, pages 511–518.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662.

Kanarios, K., Zhang, Q., and Ying, L. (2024). Cost aware best arm identification. *arXiv preprint arXiv:2402.16710*.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42.

Komiyama, J., Tsuchiya, T., and Honda, J. (2022). Minimax optimal algorithms for fixed-budget best arm identification. *Advances in Neural Information Processing Systems*, 35:10393–10404.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670.

Liu, X., Li, B., Shi, P., and Ying, L. (2021). Pond: Pessimistic-optimistic online dispatching.

Magureanu, S., Combes, R., and Proutiére, A. (2014). Lipschitz bandits: Regret lower bounds and optimal algorithms. *Proceedings of the 27th Conference on Learning Theory*.

Qin, Z., Xue, W., Zheng, L., Gan, X., Wu, H., Jin, H., and Fu, L. (2025). Cost-aware best arm identification in stochastic bandits. *ACM Transactions on Intelligent Systems and Technology*.

Shen, C. (2019). Universal best arm identification. *IEEE Transactions on Signal Processing*, 67(17):4464–4478.

Shin, J., Ramdas, A., and Rinaldo, A. (2019). On the bias, risk and consistency of sample means in multi-armed bandits. *arXiv preprint arXiv:1902.00746*.

Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. (2024). Best arm identification with fixed budget: A large deviation perspective. *Advances in Neural Information Processing Systems*, 36.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. (2012). The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556.

Zhang, Q. and Ying, L. (2024). Fast and regret optimal best arm identification: fundamental limits and low-complexity algorithms. *Advances in Neural Information Processing Systems*, 36.

Zhou, R. and Tian, C. (2022). Approximate top-m arm identification with heterogeneous reward variances. In *International Conference on Artificial Intelligence and Statistics*, pages 7483–7504. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

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

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

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

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

Supplementary Material:
Cost-Aware Optimal Pairwise Pure Exploration

A NOTATION

To improve the readability, a notation table is provided.

$\mathcal{A}$	The arm set
K	The number of arms
$\boldsymbol{\mu}$	The mean vector of the arm set
$\boldsymbol{c}$	The cost vector
S_K	The permutation group of $\mathcal{A}$
$\mathcal{G}$	The set of partition sets of the exploration task
φ	The answer function of instances
$\sigma^{\boldsymbol{\mu}}$	The index of descending order of $\boldsymbol{\mu}$ (see section 3.2)
$\mathcal{B}_{ij}$	The instances that the reward of arm-j greater than arm-i
$\mathcal{I}_m$	The indices set of some (i, j) satisfying $\mathcal{G}_m = \cap_{(i,j)\in\mathcal{I}_m}\mathcal{B}_{ij}$
δ	The confidence
$N(\boldsymbol{c})$	The set of arms associated with zero cost
$P(\boldsymbol{c})$	The set of arms associated with non-zero cost
$\mathrm{supp}(\mathcal{I})$	The support set of $\mathcal{I}$ (see section 3.3)
$N_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$	The set of zero cost arm that in the support (see section 3.3)
$P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$	The set of non-zero cost arm that in the support (see section 3.3)
$\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu})$	The alternative set of $\boldsymbol{c}, \boldsymbol{\mu}$ (see formula (2))
Δ_K	The $K - 1$ dimension probability simplex
$\Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}$	A subset of Δ_K (see formula (3))
$\boldsymbol{\omega}^*(\boldsymbol{c}, \boldsymbol{\mu})$	The optimal solution to the optimization problem
$\boldsymbol{u}^*(\boldsymbol{c}, \boldsymbol{\mu})$	The optimal pulling ratio of the positive cost arms
$\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$	The optimal pulling ratio of all arms
α	The pulling ratio of zero cost arms and non-zero cost arms
r	The parameter in α
r'	The parameter in the threshold of truncation function D_δ
γ_0	The parameter in the threshold of truncation function D_δ

B DISCUSSIONS

B.1 Broad Impact

This work introduces a general framework to study the pure-exploration problems in MAB, which broadens the scope of previous studies. With pure exploration being one of the core focuses in MAB and the theoretical nature of this work, we do not foresee major negative social impacts.

B.2 Limitations and Future Works

This work introduces a general framework to study the pure exploration problem, with a focus on the pure pairwise exploration tasks. It would be an interesting direction to further investigate the framework to its full potential beyond the pairwise requirement, leveraging solely the given partitions. Also, this work mainly considers the fixed-confidence setting, while the dual fixed-budget setting (Audibert and Bubeck, 2010; Carpentier and Locatelli, 2016; Komiyama et al., 2022; Wang et al., 2024) is also an important topic worth future studies. Moreover, this work focuses on the basic tabular scenario without assuming any relationships between arms, while the established framework and the obtained insights may also inspire additional studies in other scenarios, e.g., linear bandits (Soare et al., 2014; Jedra and Proutiere, 2020).

C LOWER BOUND

C.1 Proof of Theorem 1 of the Main Paper

Proof: Let $\varphi(\boldsymbol{\mu}) = m$ and exploration task be $\mathcal{G}$. First, we recall the following lemma from Kaufmann et al. (2016).

Lemma 2 (Lemma 1 in Kaufmann et al. (2016)) *Let v and v' be two bandit models with K arms such that for all $a \in [K]$, the distributions v_a and v'_a are mutually absolutely continuous. For any almost-surely finite stopping time σ with respect to $(\mathcal{F}_t)_t$, it holds that*

$$\sum_{a\in[K]} \mathbb{E}_v[N_a(\sigma)]\mathrm{KL}(v_a, v'_a) \geq \sup_{\mathcal{E}\in\mathcal{F}_\sigma} \hat{d}(\mathbb{P}_v(\mathcal{E}), \mathbb{P}_{v'}(\mathcal{E}))$$

where $\hat{d}(x, y) := x\log(x/y) + (1-x)\log((1-x)/(1-y))$ is the binary relative entropy, with the convention that $\hat{d}(0, 0) = \hat{d}(1, 1) = 0$.

Recall the alternative set is:

$$\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in S : \varphi(\boldsymbol{\lambda}) \neq \varphi(\boldsymbol{\mu}), \lambda_i = \mu_i, \forall i \notin P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})\}.$$

To prove Theorem 1, we can take $\mathcal{E} = \{\varphi(\boldsymbol{\mu}) = m\}$ in Lemma 2, which leads to

$$\forall\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu}), \quad \sum_{a\in\mathcal{A}} d(\mu_a, \lambda_a)\mathbb{E}[N_a(\tau_\delta)] = \sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} d(\mu_a, \lambda_a)\mathbb{E}[N_a(\tau_\delta)] \geq \mathrm{kl}(\delta, 1-\delta),$$

where the first equation is from the fact that $\mu_a = \lambda_a$ for $a \notin P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$ according to the definition of $\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu})$.

Thus, it can be obtained that

$$\begin{aligned}
\mathrm{kl}(\delta, 1-\delta) &\leq \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})} \mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)] \left(\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \frac{c_a\mathbb{E}[N_a]}{\mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)]} \frac{d(\mu_a, \lambda_a)}{c_a} \right) \\
&\leq \mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)] \sup_{\omega\in\Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})} \left(\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \omega_a \frac{d(\mu_a, \lambda_a)}{c_a} \right)
\end{aligned}$$

where the second inequality leverages the definition that $\mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)] := \sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} c_a\mathbb{E}[N_a(\tau_\delta)]$. We have

$$\mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)] \geq T^*(\boldsymbol{c}, \boldsymbol{\mu})\mathrm{kl}(\delta, 1-\delta),$$

and

$$\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)] \geq \mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)].$$

Thus,

$$\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)] \geq T^*(\boldsymbol{c}, \boldsymbol{\mu})\mathrm{kl}(\delta, 1-\delta),$$

which concludes the proof. ■

C.2 Characteristic Time and Optimal Proportions

In the following, additional properties of the lower bound are further provided. In particular, for every $\alpha \in [0,1]$, we denote

$$I_\alpha(\mu_1, \mu_2) := \alpha d(\mu_1, \alpha\mu_1 + (1-\alpha)\mu_2) + (1-\alpha) d(\mu_2, \alpha\mu_1 + (1-\alpha)\mu_2).$$

Then, the following lemma can be established, which characterizes a more explicit form of the infimum and is helpful to compute $\boldsymbol{\omega}^*$ and T^*.

Lemma 3 *For every $\omega \in \Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}$, $\varphi(\boldsymbol{\mu}) = m$ and $\mathcal{G}_m = \cap_{(i,j)\in\mathcal{I}}\mathcal{B}_{ij}$, it holds that*

$$\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})} \left(\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} \omega_a \frac{d(\mu_a, \lambda_a)}{c_a} \right) = \min \Bigg[\min_{\substack{(a,b)\in\mathcal{I}_m \\ a,b\notin N(\boldsymbol{c})}} \left(\frac{\omega_b}{c_b} + \frac{\omega_a}{c_a} \right) I_{\frac{\omega_b/c_b}{\omega_b/c_b+\omega_a/c_a}}(\mu_b, \mu_a),$$
$$\min_{\substack{(a,b)\in\mathcal{I}_m \\ a\notin N(\boldsymbol{c}),b\in N(\boldsymbol{c})}} \left(\frac{\omega_a}{c_a} d(\mu_a, \mu_b) \right), \min_{\substack{(a,b)\in\mathcal{I}_m \\ b\notin N(\boldsymbol{c}),a\in N(\boldsymbol{c})}} \left(\frac{\omega_b}{c_b} d(\mu_b, \mu_a) \right) \Bigg].$$

Proof: Without loss of generality, we assume $\mathrm{supp}(\mathcal{I}_m) = \mathcal{A}$, and then $P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}) = P(\boldsymbol{c})$ and $N_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}) = N(\boldsymbol{c})$. For $a > b, a \notin N(\boldsymbol{c}), b \in N(\boldsymbol{c})$, we can first establish that

$$\inf_{\substack{\boldsymbol{\lambda}\in S:\lambda_b=\mu_b \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)<0}} \Big(\sum_{a\in P(\boldsymbol{c})} \frac{\omega_a}{c_a} d(\mu_a, \lambda_a) \Big) = \inf_{\substack{\boldsymbol{\lambda}\in S:\lambda_b=\mu_b \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)\le 0}} \frac{\omega_a}{c_a} d(\mu_a, \lambda_a) = \frac{\omega_a}{c_a} d(\mu_a, \mu_b).$$

Similarly, for $a > b, a \in N(\boldsymbol{c}), b \notin N(\boldsymbol{c})$, we have

$$\inf_{\substack{\boldsymbol{\lambda}\in S:\lambda_a=\mu_a \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)<0}} \Big(\sum_{a\in P(\boldsymbol{c})} \frac{\omega_a}{c_a} d(\mu_a, \lambda_a) \Big) = \inf_{\substack{\boldsymbol{\lambda}\in S:\lambda_a=\mu_a \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)\le 0}} \frac{\omega_b}{c_b} d(\mu_b, \lambda_b) = \frac{\omega_b}{c_b} d(\mu_b, \mu_a).$$

Moreover, for $a > b, a \notin N(\boldsymbol{c}), b \notin N(\boldsymbol{c})$, it holds that

$$\inf_{\substack{\boldsymbol{\lambda}\in S: \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)<0}} \Big(\sum_{a\in P(\boldsymbol{c})} \frac{\omega_a}{c_a} d(\mu_a, \lambda_a) \Big) = \inf_{\substack{\boldsymbol{\lambda}\in S: \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)\le 0}} \Big(\frac{\omega_b}{c_b} d(\mu_b, \lambda_b) + \frac{\omega_a}{c_a} d(\mu_a, \lambda_a) \Big).$$

Assuming $\mu_a < \mu_b$ without loss of generality, minimizing

$$f(\lambda_b, \lambda_a) = \frac{\omega_b}{c_b} d(\mu_b, \lambda_b) + \frac{\omega_a}{c_a} d(\mu_a, \lambda_a),$$

is a convex optimization problem under the constraint $\lambda_a \ge \lambda_b$, which can be solved in closed form. In fact, the minimum is

$$\lambda_b = \lambda_a = \frac{\omega_b/c_b}{\omega_b/c_b + \omega_a/c_a}\mu_b + \frac{\omega_a/c_a}{\omega_b/c_b + \omega_a/c_a}\mu_a$$

and the optimal function value can be written as $(\frac{\omega_b}{c_b} + \frac{\omega_a}{c_a}) I_{\frac{\omega_b/c_b}{\omega_b/c_b+\omega_a/c_a}}(\mu_b, \mu_a)$.

Combining the following Lemma 4, we have

$$T^*(\boldsymbol{c},\boldsymbol{\mu})^{-1} = \sup_{\omega\in\Sigma_{P(\boldsymbol{c})}} \min_{(a,b)\in\mathcal{I}_m} \inf_{\substack{\boldsymbol{\lambda}\in S:\lambda_i=\mu_i,\forall i\in N(\boldsymbol{c}) \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)<0}} \Big(\sum_{a\in P(\boldsymbol{c})} \omega_a \frac{d(\mu_a, \lambda_a)}{c_a} \Big)$$
$$= \sup_{\omega\in\Sigma_{P(\boldsymbol{c})}} \min \Bigg[\min_{\substack{(a,b)\in\mathcal{I}_m \\ a,b\notin N(\boldsymbol{c})}} \Big(\frac{\omega_b}{c_b} + \frac{\omega_a}{c_a} \Big) I_{\frac{\omega_b/c_b}{\omega_b/c_b+\omega_a/c_a}}(\mu_b, \mu_a), \min_{\substack{(a,b)\in\mathcal{I}_m \\ a\notin N(\boldsymbol{c}),b\in N(\boldsymbol{c})}} \Big(\frac{\omega_a}{c_a} d(\mu_a, \mu_b) \Big), \min_{\substack{(a,b)\in\mathcal{I}_m \\ b\notin N(\boldsymbol{c}),a\in N(\boldsymbol{c})}} \Big(\frac{\omega_b}{c_b} d(\mu_b, \mu_a) \Big) \Bigg],$$

which concludes the proof. ■

Lemma 4 *Let $\mu \in S$ such that $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ and $\varphi(\boldsymbol{\mu}) = m$. We have*

$$\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu}) = \bigcup_{(a,b)\in\mathcal{I}_m} \{\boldsymbol{\lambda} \in S : (\lambda_a - \lambda_b)(\mu_a - \mu_b) < 0,\ \lambda_i = \mu_i, \forall i \notin P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})\}.$$

Proof: Let S_K be the permutation group of order K. Recall that

$$\mathcal{I}_m \subseteq \mathcal{H} = \{(a,b) : 1 \le b \le a \le K\}.$$

Assume that $\varphi(\boldsymbol{\mu}) = m$ and $\mathcal{G}_m = \cap_{(i,j)\in\mathcal{I}_m} \mathcal{B}_{ij}$. Consider set

$$S(\boldsymbol{\mu}) = \bigcup_{(a,b)\in\mathcal{I}_m} \mathcal{B}_{ba}.$$

Assume $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in S : \varphi(\boldsymbol{\lambda}) \neq \varphi(\boldsymbol{\mu})\}$. If $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$, there must exist $(i,j) \in \mathcal{I}$, such that $(\sigma_i^{\boldsymbol{\lambda}} - \sigma_j^{\boldsymbol{\lambda}})(\sigma_i^{\boldsymbol{\mu}} - \sigma_j^{\boldsymbol{\mu}}) < 0$. Thus

$$\mathrm{Alt}(\boldsymbol{\mu}) \subset S(\boldsymbol{\mu}).$$

Also, it is obvious that

$$\mathrm{Alt}(\boldsymbol{\mu}) \supset S(\boldsymbol{\mu}).$$

We thus have

$$\mathrm{Alt}(\boldsymbol{\mu}) = S(\boldsymbol{\mu}) = \bigcup_{(a,b)\in\mathcal{I}_m} \mathcal{B}_{ba} = \bigcup_{(a,b)\in\mathcal{I}_m} \{\boldsymbol{\lambda} \in S : (\lambda_a - \lambda_b)(\mu_a - \mu_b) < 0\}.$$

After adding cost vector $\boldsymbol{c}$ and $\mathrm{supp}(\mathcal{I}(\boldsymbol{\mu}))$, we can get

$$\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu}) = \bigcup_{(a,b)\in\mathcal{I}_m} \{\boldsymbol{\lambda} \in S : (\lambda_a - \lambda_b)(\mu_a - \mu_b) < 0,\ \lambda_i = \mu_i, \forall i \notin P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})\},$$

which concludes the proof. ■

D TRACKING RESULTS

Here, we discuss the properties of our sampling and algorithm. Recall we track the proportion of

$$\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) = \underbrace{\mathbf{1}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))}_{\text{arms in } N_{\mathcal{G}}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))} + \underbrace{(1-\alpha)\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))}_{\text{arms in } P_{\mathcal{G}}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))},$$

and when $N_{\mathcal{G}}(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) = \emptyset$, our track proportion is directly set as $\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t))$.

The proposed algorithm CAET uses a specific sample distribution $\boldsymbol{u}_\alpha(\boldsymbol{c}, \boldsymbol{\mu}) = (u_1, \dots, u_K), \alpha = \alpha(\delta)$. The following properties (Lemma 5, Proposition 3) can be obtained with their proofs extended from Lemma 7 and Proposition 9 in Garivier and Kaufmann (2016).

Lemma 5 *For all $t \ge 1$ and arm $a \in \mathcal{A}$. For any fixed $\delta > 0$ the tracking strategy with sample distribution $\boldsymbol{u}_\alpha(\boldsymbol{c}, \boldsymbol{\mu}) = (u_1, \dots, u_K)$, ensures that $N_a(t) \ge \sqrt{t + K^2} - 2K$ and*

$$\max_{1\le a\le K} \left| N_a(t) - \sum_{s=0}^{t-1} \boldsymbol{u}_a(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s)) \right| \le K(1 + \sqrt{t}).$$

When $N_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu}) = \emptyset$, our sampling strategy does not change with δ, and each arm in $\mathrm{supp}(\mathcal{I}(\boldsymbol{\mu}))$ has a positive cost.

Proposition 3 *When $N_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})=\emptyset$, for all arm $a\in\operatorname{supp}(\mathcal{I}(\boldsymbol{\mu}))$, the tracking strategy satisfies that*

$$P\left(\lim_{t\to\infty}\frac{N_a(t)}{t}=u_a\right)=1.$$

However, when $N_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})\neq\emptyset$, recall that $\boldsymbol{u}^*_\alpha=(u_1,\dots,u_K)=\mathbf{1}_\alpha(\boldsymbol{c},\boldsymbol{\mu})+\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu})$ and the optimal proportion u_a under δ is a function of δ, $u_a=u_a(\delta)$, which is no longer a constant and thus means that Proposition 3 is not sufficient. We also denote $\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu})=(u_1^*,\dots,u_K^*)$.

We need a proposition stronger than Proposition 3. Before proving Proposition 7, we introduce some lemma first.

Lemma 6 *Let $\boldsymbol{\nu}=(\nu_1,\dots,\nu_K)$ be a subgaussian random vector with $\mathbb{E}(\nu_i)=\mu_i$. $X_{a,t},\ t=1,2,\dots$, be a series i.i.d. sample from distribution ν_a. Define event*

$$\mathcal{E}_M:=\left\{\max_a\max\left(X_{a,1},\frac{1}{2}(X_{a,1}+X_{a,2}),\dots\right)\le M\right\}$$

for sufficiently large M. We have $\mathbb{P}(\mathcal{E}_M^c)\to0$, as $M\to\infty$.

Proof: Denote $\bar{\mu}_a(t)=\frac{1}{t}(X_{a,1}+\dots+X_{a,t})$ and assume ν_i are σ_i-subgaussian random variable. Let

$$\mu_{max}=\max\{\mu_1,\dots,\mu_K\}\quad\textit{and}\quad\sigma_{max}=\max\{\sigma_1,\dots,\sigma_K\}.$$

By the operation of subgaussian random variables, we have $\bar{\mu}_a(t)$ is $\sigma_i/\sqrt{t}$-subgaussian which means, for $M>\mu_{max}$, we have

$$\mathbb{P}(\bar{\mu}_a(t)>M)\le\exp\left(-\frac{t(M-\mu_a)^2}{2\sigma_a^2}\right)\le\exp\left(-\frac{t(M-\mu_{max})^2}{2\sigma_{max}^2}\right)$$

Thus, we have an union bound for $\mathbb{P}(\mathcal{E}_M)$:

$$\mathbb{P}(\mathcal{E}_M)\le\sum_a\sum_{t\ge1}\mathbb{P}(\bar{\mu}_a(t)>M)\le K\sum_{t\ge1}\exp\left(-\frac{t(M-\mu_{max})^2}{2\sigma_{max}^2}\right)\le K\frac{\exp\left(-\frac{(M-\mu_{max})^2}{2\sigma_{max}^2}\right)}{1-\exp\left(-\frac{(M-\mu_{max})^2}{2\sigma_{max}^2}\right)}$$

and

$$\lim_{M\to\infty}\exp\left(-\frac{(M-\mu_{max})^2}{2\sigma_{max}^2}\right)=0$$

leads to our result: $\mathbb{P}(\mathcal{E}_M^c)\to0$ when $M\to\infty$. ■

Lemma 7 (Continuity) *For $\boldsymbol{\mu}$ belongs to exactly one $\mathcal{G}_m\in\mathcal{G}$ and cost vector $\boldsymbol{c}$, there exists a constant δ_0 that when $\delta<\delta_0$, all of $\boldsymbol{\omega}^*(\boldsymbol{c},\boldsymbol{\mu})$, $\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu})$ and $\boldsymbol{u}_\alpha(\boldsymbol{c},\boldsymbol{\mu})$ are continuous at $(\boldsymbol{c},\boldsymbol{\mu})$.*

Proof: Recall $D_\delta([a_1,\dots,a_K])=[b_1,\dots,b_K]$ with $b_i=a_i$ if $a_i>\gamma_0\log^{-r'}(1/\delta)$ and $b_i=0$ if $a_i\le\gamma_0\log^{-r'}(1/\delta)$. We take δ_0 such that $\gamma_0\log^{-r'}(1/\delta)<\min_{i:c_i>0}\{c_i\}$. When $\delta\le\delta_0$, we have $D_\delta(\boldsymbol{c})=\boldsymbol{c}$.

Since $\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu})=G_{\boldsymbol{c}}(\boldsymbol{\omega}^*(\boldsymbol{c},\boldsymbol{\mu}))$ and $\boldsymbol{u}_\alpha(\boldsymbol{c},\boldsymbol{\mu})=\mathbf{1}(\boldsymbol{c},\boldsymbol{\mu})+\boldsymbol{\omega}^*(\boldsymbol{c},\boldsymbol{\mu})$, we only need to prove the continuity of $\mathbf{1}(\boldsymbol{c},\boldsymbol{\mu}),\boldsymbol{\omega}^*(\boldsymbol{c},\boldsymbol{\mu})$. Recall that

$$\begin{aligned}\boldsymbol{\omega}^*(\boldsymbol{c},\boldsymbol{\mu})&:=\operatorname*{arg\,max}_{\boldsymbol{\omega}\in\Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}}\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{c},\boldsymbol{\mu})}\left\{\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}\omega_a\frac{d(\mu_a,\lambda_a)}{c_a}\right\}\\&=\operatorname*{arg\,max}_{\boldsymbol{\omega}\in\Sigma_{P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})}}\min\left[\min_{\substack{(a,b)\in\mathcal{I}_m\\a,b\notin N(\boldsymbol{c})}}\left(\frac{\omega_b}{c_b}+\frac{\omega_a}{c_a}\right)I_{\frac{\omega_b/c_b}{\omega_b/c_b+\omega_a/c_a}}(\mu_b,\mu_a),\min_{\substack{(a,b)\in\mathcal{I}_m\\a\notin N(\boldsymbol{c})\\b\in N(\boldsymbol{c})}}\left(\frac{\omega_a}{c_a}d(\mu_a,\mu_b)\right),\min_{\substack{(a,b)\in\mathcal{I}_m\\b\notin N(\boldsymbol{c})\\a\in N(\boldsymbol{c})}}\left(\frac{\omega_b}{c_b}d(\mu_b,\mu_a)\right)\right]\end{aligned}\tag{10}$$

and

$$P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu}) = P(\boldsymbol{c}) \cap \operatorname{supp}(\mathcal{I}(\boldsymbol{\mu})) = \{a : a \in \mathcal{A}, c_a > 0, a \in \operatorname{supp}(\mathcal{I}(\boldsymbol{\mu}))\}.$$

Taking $\varepsilon_0 = \min\{\min_i\{c_i - \gamma_0 \log^{-r'}(1/\delta)\}, \gamma_0 \log^{-r'}(1/\delta)\}$ (note that $\delta < \delta_0$), when $|\boldsymbol{c}' - \boldsymbol{c}| < \varepsilon_0$, we have $N(D_\delta(\boldsymbol{c}')) = N(\boldsymbol{c})$ and $P(D_\delta(\boldsymbol{c}')) = P(\boldsymbol{c})$.

For $\mathcal{G}$, because $\boldsymbol{\mu}$ belongs to exactly one partition $\mathcal{G}_m$, Letting

$$\varepsilon_1 = \frac{1}{2} \min_{\substack{1 \le i,j \le K \\ \mu_i \neq \mu_j}} |\mu_i - \mu_j|$$

when $|\boldsymbol{\mu}' - \boldsymbol{\mu}| < \varepsilon_1$, from the definition of ε_1, we can have $\varphi(\boldsymbol{\mu}') = \varphi(\boldsymbol{\mu})$.

Form the above discussion, for all $(\boldsymbol{c}', \boldsymbol{\mu}')$ satisfying $|\boldsymbol{\mu}' - \boldsymbol{\mu}| < \varepsilon_1$ and $|\boldsymbol{c}' - \boldsymbol{c}| < \varepsilon_0$, there is $P_{\mathcal{G}}(\boldsymbol{c}', \boldsymbol{\mu}') = P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$ and we can discuss the continuity in this neighborhood of $(\boldsymbol{c}, \boldsymbol{\mu})$ named it as $\mathcal{N}$. In this neighborhood, $\mathcal{I}(\boldsymbol{\mu}') \equiv (\boldsymbol{\mu})$ and $N(D_\delta(c')) \equiv N(\boldsymbol{c})$, which suggests that $\boldsymbol{\omega}^*(\boldsymbol{c}, \boldsymbol{\mu})$ is made of some continuous function and some math operation (min, arg max) by Equation (10). Thus, $\boldsymbol{\omega}^*(\boldsymbol{c}, \boldsymbol{\mu})$ is continuous at $(\boldsymbol{c}, \boldsymbol{\mu})$ when $\delta < \delta_0$. For $\mathbf{1}_\alpha(\boldsymbol{c}, \boldsymbol{\mu})$, since in neighborhood $\mathcal{N}$, $P_{\mathcal{G}}(\boldsymbol{c}', \boldsymbol{\mu}') = P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$, so $\mathbf{1}_\alpha(\boldsymbol{c}, \boldsymbol{\mu}) = \mathbf{1}_\alpha(\boldsymbol{c}', \boldsymbol{\mu}')$ for $(\boldsymbol{c}', \boldsymbol{\mu}') \in \mathcal{N}$. Thus, $\mathbf{1}_\alpha(\boldsymbol{c}, \boldsymbol{\mu})$ is also continuous and we conclude the proof. ■

Definition 2 *For bandits model $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, we call a sampling rule $\mathcal{R}$ **stable** if there exist constants $0 < C_1$ and $C_2 < 1$ that*

$$\lim_{t \to \infty} \mathbb{P}\left(C_1 < \frac{N_a(t)}{t} < C_2\right) = 1, \quad \forall a = 1, \ldots, K.$$

Recall the stopping rule of CAET algorithm

$$\tau_\delta = \inf\{t \in N : \exists \mathcal{G}_m, \forall (a,b) \in \mathcal{I}_m, Z_{a,b}(t) > \beta(t,\delta)\}, \tag{11}$$

and we now present some properties regarding τ_δ.

Proposition 4 *For bandits model $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, any stable sampling rule will ensure an almost sure finite stopping time τ_δ under the stopping rule given in Equation* (11) *with $\beta(t,\delta) = \log(Ct^\theta/\delta), \theta \in [1, e/2]$, i.e.,*

$$\mathbb{P}(\tau_\delta < \infty) = 1.$$

Proof: Since the exploration task is pairwise, and contains only finite pairwise tests, we only need to consider the two-arm situation and Proposition 5 is enough. ■

Proposition 5 *For the two-arm bandit model $\boldsymbol{\mu} = (\mu_1, \mu_2)$, any stable sampling rule will ensure an almost sure finite stopping time τ_δ under the stopping rule given in Equation* (11) *with $\beta(t,\delta) = \log(Ct^\theta/\delta), \theta \in [1, e/2]$, i.e.,*

$$\mathbb{P}(\tau_\delta < \infty) = 1.$$

Proof: Recall when $\hat{\mu}_1(t) > \hat{\mu}_2(t)$,

$$Z_{1,2}(t) = N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_{1,2}(t)) + N_2(t) d(\hat{\mu}_2(t), \hat{\mu}_{1,2}(t)).$$

Let $\mathcal{E} = \{\forall a \in \{1,2\}, \hat{\mu}_a(t) \underset{t \to \infty}{\to} \mu_a\}$. From the assumption on the sampling strategy and the Law of Large Numbers, $\mathbb{P}(\mathcal{E}) = 1$. When t goes to infinity, both $d(\hat{\mu}_1(t), \hat{\mu}_{1,2}(t))$ and $d(\hat{\mu}_2(t), \hat{\mu}_{1,2}(t))$ are lower bounded by some constant $K(\boldsymbol{\mu})$ where

$$K(\boldsymbol{\mu}) = \min\left\{d(\mu_1, C_2\mu_1 + (1 - C_2)\mu_2), d(\mu_2, C_2\mu_2 + (1 - C_2)\mu_1)\right\}.$$

Therefore, for all $\epsilon > 0$ there exists t_1 such that for all $t > t_1$

$$d(\hat{\mu}_1(t), \hat{\mu}_{1,2}(t)) > \frac{1}{1+\epsilon} K(\boldsymbol{\mu}),$$
$$d(\hat{\mu}_2(t), \hat{\mu}_{1,2}(t)) > \frac{1}{1+\epsilon} K(\boldsymbol{\mu}).$$

Hence, for $t > t_1$,

$$Z_{1,2}(t) > 2tC_1 \frac{1}{1+\epsilon} K(\boldsymbol{\mu}).$$

Consequently,

$$\tau_\delta = \inf\{t \in \mathbb{N} : Z_{1,2}(t) \geq \beta(t,\delta)\} \leq t_1 \vee \inf\{t \in \mathbb{N} : 2tC_1(1+\epsilon)^{-1}K(\boldsymbol{\mu}) \geq \log(Ct^\alpha/\delta)\}.$$

Using Lemma 8 below, it follows that on $\mathcal{E}$, as $\alpha \in [1, e/2]$,

$$\tau_\delta \leq t_1 \vee \alpha(1+\epsilon)C_1^{-1}K(\boldsymbol{\mu})^{-1} \left[\log\left(\frac{Ce(1+\epsilon)^\alpha}{\delta(C_1K(\boldsymbol{\mu}))^\alpha} \right) + \log\log\left(\frac{C(1+\epsilon)^\alpha}{\delta(C_1K(\boldsymbol{\mu}))^\alpha} \right) \right].$$

Thus τ_δ is finite on $\mathcal{E}$, which concludes the proof. ■

Lemma 8 *For every $\alpha \in [1, e/2]$, for any two constant $\kappa_1, \kappa_2 > 0$, there exists*

$$x = \frac{\alpha}{\kappa_1} \left[\log\left(\frac{\kappa_2 e}{\kappa_1^\alpha} \right) + \log\log\left(\frac{\kappa_2}{\kappa_1^\alpha} \right) \right]$$

such that $\kappa_1 x \geq \log(\kappa_2 x^\alpha)$.

Proof: This lemma has been stated in Garivier and Kaufmann (2016) which can be checked directly. It can also be seen as a by-product of well-known bounds on the Lambert W function. ■

Lemma 9 *Let $\theta \in [1, e/2]$ and assume $\mathrm{supp}(\mathcal{I}(\boldsymbol{\mu})) = \mathcal{A}$. For any exploration task $\mathcal{G}$ in bandit model $\boldsymbol{\mu}$, denote $\{\mathcal{R}_\delta\}$ as a series of stable sampling strategies. Each sampling rule $\mathcal{R}_\delta$, with the stopping rule given in Equation* (11) *under threshold $\beta(t,\delta) = \log(Ct^\theta/\delta)$, has a stopping time $\tau_\delta = \tau_\delta(\mathcal{R}_\delta)$ and respective pulling time $N_a^\delta(\tau_\delta)$. Then $\{\tau_\delta\}_\delta$ is a sequence of random variables that almost surely satisfy*

$$\mathbb{P}\left(\liminf_{\delta \to 0} \frac{N_a^\delta(\tau_\delta)}{\log(1/\delta)} > K(\boldsymbol{\mu}) \right) = 1, \qquad a \in \mathcal{A},$$

where $K(\boldsymbol{\mu})$ is a constant of $\boldsymbol{\mu}$.

Proof: Define event $\Omega_0 = \{\bar{\mu}_a(t) \underset{t\to\infty}{\rightarrow} \mu_a, \forall a\} \cap \{\bar{c}_a(t) \underset{t\to\infty}{\rightarrow} c_a, \forall a\}$. From the Law of Large Numbers, $\mathbb{P}(\Omega_0) = 1$. Considering the event $\mathcal{E}_m \cap \Omega_0$, we will prove the convergence conditioned on this event and then let M go to infinity to conclude the proof (we use the fact of Lemma 6: $\lim_{M\to\infty} \mathbb{P}(\mathcal{E}_M) = 1$).

Recall the stopping rule is

$$\tau_\delta = \inf\{t \in N : \exists \mathcal{G}_m, \forall (a,b) \in \mathcal{I}_m, Z_{a,b}(t) > \beta(t,\delta)\},$$

and when $\hat{\mu}_a(t) > \hat{\mu}_b(t)$,

$$Z_{a,b}(t) = N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{a,b}(t)) + N_b(t) d(\hat{\mu}_b(t), \hat{\mu}_{a,b}(t))$$

where

$$\hat{\mu}_{a,b}(t) = \frac{N_a(t)}{N_a(t)+N_b(t)} \hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t)+N_b(t)} \hat{\mu}_b(t).$$

First, we prove $\mathbb{P}(\lim_{\delta\to 0}\tau_\delta = \infty) = 1$:

$$\begin{aligned}\log(C/\delta) <& \log(C{\tau_\delta}^\theta/\delta)\\ <& N_a^\delta(\tau_\delta)d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) + N_b^\delta(\tau_\delta)d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))\\ <& \tau_\delta[d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) + d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))].\end{aligned}$$

Since conditioned on event $\mathcal{E}_M$, both $d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))$ and $d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))$ are bounded, we get

$$\mathbb{P}(\lim_{\delta\to 0}\tau_\delta = \infty|\mathcal{E}_M) = 1$$

and let M approach to infinity, by Lemma 6, we can get the desire result:

$$\mathbb{P}(\lim_{\delta\to 0}\tau_\delta = \infty) = 1.$$

For any $b \in \text{supp}(\mathcal{I}(\boldsymbol{\mu})) = \mathcal{A}$, there exists a, such that $(a, b) \in \mathcal{I}(\boldsymbol{\mu})$ or $(b, a) \in \mathcal{I}(\boldsymbol{\mu})$. We assume $(a, b) \in \mathcal{I}(\boldsymbol{\mu})$. Then, τ_δ satisfies

$$N_a^\delta(\tau_\delta)d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) + N_b^\delta(\tau_\delta)d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) > \log(C{\tau_\delta}^\theta/\delta).$$

Assuming $N_a^\delta(\tau_\delta) > N_b^\delta(\tau_\delta)$, we first consider the following limits

$$\lim_{t=x/y\to\infty} x\cdot d\left(\mu_a, \frac{x\mu_a + y\mu_b}{x+y}\right)\bigg/ y = \lim_{t\to\infty}\frac{d\left(\mu_a, \frac{t\mu_a+\mu_b}{t+1}\right)}{1/t}.$$

By the L'Hospital rule, we only need to calculate

$$\begin{aligned}&\lim_{t\to\infty}\frac{\frac{d}{dt}d\left(\mu_a, \frac{t\mu_a+\mu_b}{t+1}\right)}{-1/t^2} = \lim_{t\to\infty} -t^2\frac{d}{dt}d\left(\mu_a, \frac{t\mu_a+\mu_b}{t+1}\right)\\ =&\lim_{t\to\infty} -t^2\frac{(\mu_b-\mu_a)}{t+1}\frac{1}{b''(b^{-1}(y_0))}\frac{(\mu_a-\mu_b)}{(t+1)^2}\\ =&\lim_{t\to\infty}\frac{(\mu_a-\mu_b)^2}{b''(b^{-1}(y_0))}\frac{t^2}{(t+1)^3}\\ =&0\end{aligned}$$

where $\frac{d}{dx}d(x, y) = (y - x)/b''(b^{-1}(y))$ and

$$y_0 = \frac{t\mu_a + \mu_b}{t+1} \in [\mu_a, \mu_b]$$

is bounded. So, we have

$$\lim_{t=x/y\to\infty} x\cdot d\left(\mu_1, \frac{x\mu_a + y\mu_b}{x+y}\right)\bigg/ y = 0$$

which shows there exists $N_0 \in \mathbb{N}$ such that when $x/y > N_0$, we have

$$x\cdot d\left(\mu_a, \frac{x\mu_a + y\mu_b}{x+y}\right) < y.$$

Now, we continue our proof. If $N_a^\delta(\tau_\delta)/N_b^\delta(\tau_\delta) > N_0$, we have

$$N_a^\delta(\tau_\delta)d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) < N_b^\delta(\tau_\delta).$$

Thus,

$$\log(C/\delta) < \log(C{\tau_\delta}^\theta/\delta)$$

$$
\begin{aligned}
&< N_a^\delta(\tau_\delta) d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) + N_b^\delta(\tau_\delta) d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) \\
&< N_b^\delta(\tau_\delta) + N_b^\delta(\tau_\delta) d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) \\
&< N_b^\delta(\tau_\delta)(1 + d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))).
\end{aligned}
$$

When $N_a^\delta(\tau_\delta)/N_b^\delta(\tau_\delta) < N_0$, we have

$$
\begin{aligned}
\log(C/\delta) &< \log(C{\tau_\delta}^\theta/\delta) \\
&< N_a^\delta(\tau_\delta) d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) + N_b^\delta(\tau_\delta) d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) \\
&< N_b^\delta(\tau_\delta) \big[N_0 d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) + d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) \big].
\end{aligned}
$$

Thus

$$
\begin{aligned}
\log(C/\delta) < N_b^\delta(\tau_\delta) \cdot \max \big\{ &(1 + d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))), \\
&(N_0 d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta)) + d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))) \big\}.
\end{aligned} \tag{12}
$$

Since $\mathbb{P}(\lim_{\delta \to 0} \tau_\delta = \infty) = 1$ and Lemma 5 ensures that $N^\delta(\tau_\delta) \geq \sqrt{\tau_\delta + K^2} - 2K$, we can have that, for all arm $a \in \mathcal{A}$,

$$
\mathbb{P}(\liminf_{\delta \to 0} N_a^\delta(\tau_\delta) = \infty) = 1, \quad a \in \mathcal{A}
$$

Therefore, by the Law of Large Numbers, both $d(\hat{\mu}_a(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))$ and $d(\hat{\mu}_b(\tau_\delta), \hat{\mu}_{a,b}(\tau_\delta))$ are bounded with probability 1. Thus, by Equation (12), we have

$$
\mathbb{P}\left(\liminf_{\delta \to 0} \frac{N_b^\delta(\tau_\delta)}{\log(1/\delta)} > K(\boldsymbol{\mu}) \right) = 1
$$

for some constant $K(\boldsymbol{\mu})$. Repeating the same process for the other arms, we have

$$
\mathbb{P}\left(\liminf_{\delta \to 0} \frac{N_a^\delta(\tau_\delta)}{\log(1/\delta)} > \hat{K}(\boldsymbol{\mu}) \right) = 1, \quad a \in \mathcal{A},
$$

which concludes the proof. ■

Before proving the main proposition, we introduce a concentration lemma to handle the bias of the sample means during adaptive sampling and adaptive stopping.

Lemma 10 *Consider an adaptive sampling algorithm and stopping rule. For a fixed arm $k \in \mathcal{A}$ with a finite $2p$-norm, where $p > 1$, and any random time τ such that $N_k(\tau) \geq 3$ almost surely, it holds that, for any $\varsigma > 0$,*

$$
\mathbb{P}\left(\frac{N_k(\tau)}{\log N_k(\tau)} \left(\frac{\hat{\mu}_k(\tau) - \mu_k}{q_k} \right)^2 \geq \varsigma \right) \leq \frac{C_p}{\varsigma^p}
$$

where C_p is a constant depending only on p, and q_k is a constant depending only on μ_k.

The proof of this lemma has been given in Lemma 4.4 in Shin et al. (2019). For any $\delta > 0$, our sampling rule $\mathcal{R}_\delta$ satisfies $N_k(t) \geq \sqrt{t + K^2} - 2K, \ \ \forall k \in \mathcal{A}$. Thus, when $\tau = \tau'_\delta$ is large enough, we have

$$
\frac{N_k(\tau'_\delta)}{\log N_k(\tau'_\delta)} > N_k(\tau'_\delta)^{2/3} > {\tau'_\delta}^{1/4},
$$

and

$$
\mathbb{P}\left({\tau'_\delta}^{1/4} \left(\frac{\hat{\mu}_k(\tau'_\delta) - \mu_k}{q_k} \right)^2 \geq \varsigma \right) \leq \frac{C_p}{\varsigma^p}.
$$

We take $\tau'_\delta = \log^{r_1}(1/\delta), r_1 < 1$. When δ is small enough, we will have $\tau'_\delta \leq \tau_\delta$. Thus, we have

$$
\mathbb{P}\left(\hat{\mu}_k(\tau'_\delta) - \mu_k \geq q_k \varsigma^{1/2} (\tau'_\delta)^{-1/8} \right) \leq \frac{C_p}{\varsigma^p}. \tag{13}
$$

The above discussion leads to the following lemma:

Proposition 6 *For the series of sampling rules $\{\mathcal{R}_\delta\}_\delta$ in the CAET algorithm, we can precisely determine the arm with zero cost before a time on the order of $\tau'_\delta = \log^{r_1}(1/\delta) = o(\log(1/\delta)) = o(\tau_\delta(1-\alpha))$, i.e.,*

$$\mathbb{P}\left(\lim_{\delta\to 0} N(\boldsymbol{c}_\delta(\tau'_\delta)) = N(\boldsymbol{c})\right) = 1.$$

Proof: Recall $\boldsymbol{c}_\delta(t) = D_\delta(\hat{\boldsymbol{c}}(t))$ and $D_\delta(\cdot)$ is a truncation function with threshold $\gamma_0 \log^{-r'}(1/\delta)$. When we take $\varsigma = \gamma_0^2 \log^{r_1/4-2r'}(1/\delta)/q_k^2$ in Equation (13), it becomes:

$$\mathbb{P}\left(\hat{\mu}_k(\tau'_\delta) - \mu_k \geq \gamma_0 \log^{-r'}(1/\delta)\right) \leq \frac{C_p}{\varsigma^p}.$$

And for any $r' < 1/8$, we can take $r_1 = 1/2 + 4r' < 1$. Thus, $2r' < r_1/4$ leads to $\lim_{\delta\to 0} \varsigma = \infty$ which concludes the proof by letting δ goes to zero. ■

In the following, we introduce the main proposition which shows the desired tracking result.

Proposition 7 *For $\alpha = 1 - \log^{-r}(1/\delta)$ with $0 < r < 1/2$, the sampling rule and stopping rule given in Equation* (11) *with stopping time τ_δ satisfies*

$$\mathbb{P}\left(\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{(1-\alpha)\tau_\delta} = u_a^*\right) = 1, \quad a \in P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$$

and

$$\mathbb{P}\left(\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{\tau_\delta} = \frac{1}{|N_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})|}\right) = 1, \quad a \in N_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$$

and for arms out of $\mathrm{supp}(\mathcal{I}(\boldsymbol{\mu}))$*,*

$$\mathbb{P}\left(\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{(1-\alpha)\tau_\delta} = 0\right) = 1, \quad a \notin \mathrm{supp}(\boldsymbol{\mu}).$$

Proof sketch: To prove this proposition, we focus on a special event $\Omega_0 = \{\bar{\mu}_a(t) \underset{t\to\infty}{\to} \mu_a, \forall a\} \cap \{\bar{c}_a(t) \underset{t\to\infty}{\to} c_a, \forall a\} \cap \Omega_1 \cap \Omega_2$ and $\mathbb{P}(\Omega_0) = 1$. In this event, we can eliminate the effect of the force exploration during the sampling phase by Lemma 5, i.e.

$$\max_{1\leq a\leq K} \left| N_a(\tau_\delta) - \sum_{s=0}^{\tau_\delta - 1} \boldsymbol{u}_a(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s)) \right| \leq K(1 + \sqrt{\tau_\delta}).$$

We can prove:

$$\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{\tau_\delta(1-\alpha)} = \lim_{\delta\to 0} \sum_{s=0}^{\tau_\delta - 1} \boldsymbol{u}_a(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))/(\tau_\delta(1-\alpha)).$$

We successfully change the original estimator u_a^ϵ to a more familiar estimator u_a and the last is without force exploration. After that, we need to calculate

$$\lim_{\delta\to 0} \sum_{s=0}^{\tau_\delta - 1} \boldsymbol{u}_a(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))/(\tau_\delta(1-\alpha)).$$

Proposition 6 shows a sub-linear commitment time of committing to $N(\boldsymbol{c})$, where the sub-linearity is with respect to the order of pulling times of arms in $P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$. It means the total inaccurate time has a sub-linear order. With that, we introduce a result of sequential convergence in Lemma 11 (the details of this lemma will be stated later):

$$S_\delta = b_1 + b_2 + \cdots + b_{m_\delta} + f(\delta)(a_1 + a_2 + \cdots + a_{n_\delta - m_\delta}) \xrightarrow[\lim_{n\to\infty} a_n = a]{\text{some conditions}} \lim_{\delta\to 0} S_\delta/(n_\delta f(\delta)) = a$$

and we can get the desired result. ■

Proof: The critical point is proving the case of arm $a \in P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$. First, we consider arm $a \in P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$ and we need to prove

$$\mathbb{P}\left(\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{(1-\alpha)\tau_\delta} = u_a^*\right) = 1.$$

Denote event $\Omega_1 = \{\lim_{\delta\to 0} N(\boldsymbol{c}_\delta(\tau_\delta')) = N(\boldsymbol{c})\}$ and from Proposition 6, $\mathbb{P}(\Omega_1) = 1$.

By Lemma 9, there exists a constant $K(\boldsymbol{\mu})$ that the cumulative pulling time of arms in $P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$ before algorithm stop satisfies

$$\mathbb{P}\left(\lim_{\delta\to 0} \frac{\sum_{a\in P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})} N_a(\tau_\delta)}{\log(1/\delta)} > K(\boldsymbol{\mu})\right) = 1,$$

and, scaling with ratio $1-\alpha = \log^r(1/\delta)$, the total stopping time τ_δ satisfies

$$\mathbb{P}\left(\lim_{\delta\to 0} \frac{\tau_\delta}{\log^{1+r}(1/\delta)} > K(\boldsymbol{\mu})\right) = 1,$$

which we denoted as event Ω_2. Define event $\Omega_0 = \{\bar{\mu}_a(t) \underset{t\to\infty}{\to} \mu_a, \forall a\} \cap \{\bar{c}_a(t) \underset{t\to\infty}{\to} c_a, \forall a\} \cap \Omega_1 \cap \Omega_2$ and from the Law of Large Numbers and previous discussions, we have $\mathbb{P}(\Omega_0) = 1$.

Recall that

$$\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)) = \mathbf{1}_\alpha(\boldsymbol{c}_\delta(t), \boldsymbol{\mu}(t)) + (1-\alpha)\boldsymbol{u}^*(\boldsymbol{c}_\delta(t), \hat{\boldsymbol{\mu}}(t)).$$

Using Lemma 5, we have

$$\max_{1\le a\le K} \left| N_a(\tau_\delta) - \sum_{s=0}^{\tau_\delta - 1} \boldsymbol{u}_a(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s)) \right| \le K(1+\sqrt{\tau_\delta}).$$

Thus,

$$\max_{1\le a\le K} \left| \frac{N_a(\tau_\delta)}{\tau_\delta(1-\alpha)} - \sum_{s=0}^{\tau_\delta - 1} \boldsymbol{u}_a(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))/(\tau_\delta(1-\alpha)) \right| \le K\frac{1+\sqrt{\tau_\delta}}{\tau_\delta(1-\alpha)}.$$

From the assumption of Ω_2, $\tau_\delta = O(\log^{1+r}(1/\delta))$ and we have

$$\frac{1+\sqrt{\tau_\delta}}{\tau_\delta(1-\alpha)} \le O\left(\frac{\log^{(1+r)/2}(1/\delta)}{\log(1/\delta)}\right) = O\left(\log^{(r-1)/2}(1/\delta)\right).$$

Also $r < 1$ leads to

$$\lim_{\delta\to 0} K\frac{1+\sqrt{\tau_\delta}}{\tau_\delta(1-\alpha)} = 0.$$

We get

$$\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{\tau_\delta(1-\alpha)} = \lim_{\delta\to 0} \sum_{s=0}^{\tau_\delta - 1} \boldsymbol{u}_a(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))/(\tau_\delta(1-\alpha)).$$

For $a \in P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$, denotes $\boldsymbol{u}_a^*(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s)) = b_s$ be the a-th component of $\boldsymbol{u}^*(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))$. On event Ω_0, we have $\lim_{s\to\infty} b_s = u_a^*$ and $u_{\alpha,a}(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s)) = (1-\alpha)b_s$ where $u_{\alpha,a}(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))$ is a-th component of $\boldsymbol{u}_\alpha(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))$.

Before time τ'_δ, the estimation of $N(\boldsymbol{c})$ and $P(\boldsymbol{c})$ might be inaccurate which will have influence on $u_{\alpha,a}(\boldsymbol{c}_\delta(s), \hat{\boldsymbol{\mu}}(s))$ and we denote it as b'_s. Then, we have

$$\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{(1-\alpha)\tau_\delta} = u_a^* \Leftrightarrow \lim_{\delta\to 0} \frac{b'_1 + \cdots + b'_{\tau'_\delta} + (1-\alpha)(b_{\tau'_\delta+1} + \cdots + b_{\tau_\delta})}{\tau_\delta(1-\alpha)} = u_a^*$$

and by $0 < r < 1$, we have

$$\frac{\tau'_\delta}{\tau_\delta(1-\alpha)} = O(\log^{r_1-1}(1/\delta)), \quad \frac{\tau'_\delta}{\tau_\delta} = O(\log^{r_1-1-r}(1/\delta)).$$

Thus, applying Lemma 11, we can get, for arm $a \in P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$

$$\mathbb{P}\left(\lim_{\delta\to 0} \frac{N_a(\tau_\delta)}{(1-\alpha)\tau_\delta} = u_a^*\right) = 1, \quad \forall a \in P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu}).$$

For arms in $N_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu})$ and arms out of $\text{supp}(\mathcal{I}(\boldsymbol{\mu}))$ we can repeat the same process and we can conclude the proof. ■

Lemma 11 *Let $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ be two bounded sequences with $\lim_{n\to\infty} a_n = a$. For $\delta \in (0,1)$, n_δ, m_δ are functions of $(0,1) \to \mathbb{N}$ with $n_\delta > m_\delta$ for each δ and $f(\delta)$ is a continuous function with $f(\delta) \neq 0$. Define S_δ as*

$$S_\delta = b_1 + b_2 + \cdots + b_{m_\delta} + f(\delta)(a_1 + a_2 + \cdots + a_{n_\delta - m_\delta}).$$

When m_δ and n_δ satisfy $\lim_{\delta\to 0} m_\delta/n_\delta = 0$ and $\lim_{\delta\to 0} m_\delta/(f(\delta)n_\delta) = 0$, we have $\lim_{\delta\to 0} S_\delta/(n_\delta f(\delta)) = a$.

Proof: Assume $|b_n| \leq B$ and we have

$$\frac{S_\delta}{n_\delta f(\delta)} = \frac{b_1 + \cdots + b_{m_\delta}}{n_\delta f(\delta)} + \frac{a_1 + \cdots + a_{n_\delta - m_\delta}}{n_\delta}.$$

For the first term

$$\frac{b_1 + \cdots + b_{m_\delta}}{n_\delta f(\delta)} < B \cdot \frac{m_\delta}{n_\delta f(\delta)} \to 0,$$

which indicates that

$$\lim_{\delta\to 0} \frac{b_1 + \cdots + b_{m_\delta}}{n_\delta f(\delta)} = 0.$$

Since $m_\delta/n_\delta \to 0$ and m_δ is always no less than 1, we can get $n_\delta \to \infty$ and also

$$\lim_{\delta\to 0}(n_\delta - m_\delta) = \lim_{\delta\to 0} n_\delta \cdot (1 - \frac{m_\delta}{n_\delta}) = \infty.$$

So, it holds that

$$\begin{aligned}\lim_{\delta\to 0} \frac{a_1 + \cdots + a_{n_\delta - m_\delta}}{n_\delta} &= \lim_{\delta\to 0} \frac{a_1 + \cdots + a_{n_\delta - m_\delta}}{n_\delta - m_\delta} \cdot \frac{n_\delta - m_\delta}{n_\delta} \\ &= \lim_{\delta\to 0} \frac{a_1 + \cdots + a_{n_\delta - m_\delta}}{n_\delta - m_\delta}.\end{aligned}$$

Since $n_\delta - m_\delta \to \infty$, we only need to proof the following form

$$\lim_{n\to\infty} \frac{a_1 + a_2 + \cdots + a_n}{n} = a.$$

Assume $|a_n - a| < A, \forall n$, and for $\forall \varepsilon > 0$, $\exists N_1$, s.t., $\forall n > N_1, |a_n - a| < \varepsilon/2$. When $n > N = \frac{2N_1 A}{\varepsilon}$, we have

$$\begin{aligned}\left|\frac{a_1 + a_2 + \cdots + a_n}{n} - a\right| &\leq \left|\frac{a_1 - a}{n} + \cdots + \frac{a_n - a}{n}\right| \\ &\leq \frac{|a_1 - a|}{n} + \cdots + \frac{|a_n - a|}{n} \\ &< \frac{N_1 A}{n} + \frac{\varepsilon(n - N_1)}{2n} < \varepsilon,\end{aligned}$$

which concludes the proof of Lemma 11. ■

E PROOFS OF SECTION 6

E.1 Proof of Proposition 1

Proof: Denote $\varphi(\boldsymbol{\mu}) = m$ and $\mathcal{G}_m = \cap_{(i,j)\in\mathcal{I}_m} B_{ij}$. Let $T_{a,b} := \inf\{t \in \mathbb{N} : Z_{a,b}(t) > \beta(t,\delta)\}$, one has

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty, &\varphi(\hat{\boldsymbol{\mu}}(\tau_\delta)) \neq \varphi(\boldsymbol{\mu}))\\ &\leq \mathbb{P}_{\boldsymbol{\mu}}(\exists (a,b) \in \mathcal{I}_m, \exists t \in \mathbb{N} : Z_{b,a} > \beta(t,\delta))\\ &\leq \sum_{(a,b)\in\mathcal{I}_m} \mathbb{P}_{\mu}(T_{b,a} < \infty).\end{aligned}$$

As shown in Garivier and Kaufmann (2016), we can have the following lemma.

Lemma 12 *For any a, b such that $\mu_a < \mu_b$, let $T_{a,b} := \inf\{t \in \mathbb{N} : Z_{a,b}(t) > \beta(t,\delta)\}$ with $\beta(t,\delta) = \log(2tK(K-1)/\delta)$, we have*

$$\mathbb{P}_{\mu}(T_{a,b} < \infty) \leq \frac{\delta}{K(K-1)}.$$

Thus, using Lemma 12, we have

$$\mathbb{P}_{\mu}(\tau_\delta < \infty, \varphi(\hat{\boldsymbol{\mu}}(\tau_\delta)) \neq \varphi(\boldsymbol{\mu})) \leq |\mathcal{I}_m| \cdot \frac{\delta}{K(K-1)} \leq \delta.$$

which completes the proof. ■

E.2 Proof of Proposition 2

Proof: It can be verified that if $x > y$,

$$I_\alpha(x,y) = \inf_{x' < y'} \left[\alpha d(x,x') + (1-\alpha) d(y,y')\right].$$

For every a, b that $\mu_a < \mu_b$ and $\hat{\mu}_a(t) > \hat{\mu}_a(t)$, we have

$$\begin{aligned}Z_{a,b}(t) &= (N_a(t) + N_b(t)) I_{\frac{N_a(t)}{N_a(t)+N_b(t)}}(\hat{\mu}_a(t), \hat{\mu}_b(t))\\ &= \inf_{\mu_a' < \mu_b'} N_a(t) d(\hat{\mu}_a(t), \mu_a') + N_b(t) d(\hat{\mu}_b(t), \mu_b')\\ &\leq N_a(t) d(\hat{\mu}_a(t), \mu_a) + N_b(t) d(\hat{\mu}_b(t), \mu_b).\end{aligned}$$

For $\mu_a > \mu_b$, repeating the previous discussion, we have the same formula:

$$Z_{b,a}(t) \leq N_a(t) d(\hat{\mu}_a(t), \mu_a) + N_b(t) d(\hat{\mu}_b(t), \mu_b).$$

Assuming $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$ and $\varphi(\boldsymbol{\mu}) = m$, $\mathcal{G}_m = \cap_{(i,j)\in\mathcal{I}_m} B_{i,j}$, we have

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta \leq \infty, \varphi(\hat{\boldsymbol{\mu}}(\tau_\delta)) \neq \varphi(\boldsymbol{\mu}))\\ \leq &\mathbb{P}_{\boldsymbol{\mu}}(\exists (a,b) \in \mathcal{I}_m, \exists t \in \mathbb{N} : \hat{\mu}_b(t) > \hat{\mu}_a(t), Z_{b,a} > \beta(t,\delta))\\ \leq &\mathbb{P}_{\boldsymbol{\mu}}(\exists (a,b) \in \mathcal{I}_m, \exists t \in \mathbb{N} : N_a(t) d(\hat{\mu}_a(t), \mu_a) + N_b(t) d(\hat{\mu}_b(t), \mu_b) \geq \beta(t,\delta))\\ \leq &\mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq \beta(t,\delta)\right)\\ \leq &\sum_{t=1}^{\infty} e^{K+1} \left(\frac{\beta(t,\delta)^2 \log(t)}{K}\right)^K e^{-\beta(t,\delta)}.\end{aligned}$$

The last inequality follows from Magureanu et al. (2014) whose result can straightforwardly generalize to a one-parameter exponential family. With a rate form $\beta(t,\delta)=\log(Ct^{\theta}/\delta)$, we need to prove that for any $\delta\in(0,1)$, there exists $C=O(\text{polylog}(1/\delta))$ in $\beta(t,\delta)=\log(Ct^{\theta}/\delta)$ such that

$$\sum_{t=1}^{\infty} e^{K+1}\left(\frac{\beta(t,\delta)^2\log(t)}{K}\right)^{K} e^{-\beta(t,\delta)}\leq\delta.$$

It is equivalent to prove that for any $\delta\in(0,1)$ there exists a function of δ, $C=C(\delta)=O(\text{polylog}(1/\delta))$, such that

$$\sup_{\delta\in(0,1)}\sum_{t=1}^{\infty}\frac{e^{K+1}}{K^K}\frac{(\log^2(Ct^{\theta}/\delta)\log t)^K}{Ct^{\theta}}\leq 1.$$

We can prove this by taking $C=C_0(e+\log^{2K+1}(1/\delta))=C_0f_0(\delta)$ with a sufficiently large constant C_0.

For any $t\in\mathbb{N},\delta\in(0,1)$, we have

$$f_0(\delta)t^{\theta}/\delta>e.$$

Take sufficiently large C_0 that

$$\log(C_0)\geq 1+\frac{1}{\log(f_0(\delta)t^{\theta}/\delta)-1},$$

which implies

$$\log(C_0f_0(\delta)t^{\theta}/\delta)=\log(C_0)+\log(f_0(\delta)t^{\theta}/\delta)\leq\log(C_0)\log(f_0(\delta)t^{\theta}/\delta). \tag{14}$$

Similarly, we have

$$\log(f_0(\delta)t^{\theta}/\delta)\leq\log(f_0(\delta)/\delta)\log(t^{\theta}). \tag{15}$$

It then follows that

$$\begin{aligned}&\sum_{t=1}^{\infty}\frac{e^{K+1}}{K^K}\frac{(\log^2(C_0f_0(\delta)t^{\theta}/\delta)\log t)^K}{C_0f_0(\delta)t^{\theta}}\\ &\leq\frac{e^{K+1}}{K^K}\sum_{t=1}^{\infty}\frac{(\log^2(C_0)\cdot\log^2(f_0(\delta)t^{\theta}/\delta)\log t)^K}{C_0f_0(\delta)t^{\theta}}\\ &\leq\frac{\log^{2K}(C_0)}{C_0}\frac{e^{K+1}}{K^K}\frac{\log^{2K}(f_0(\delta)/\delta)}{f_0(\delta)}\sum_{t=1}^{\infty}\frac{\log^{2K}(t^{\theta})\log^K(t)}{t^{\theta}}.\end{aligned}$$

Note that the series $\sum_{t=1}^{\infty}\frac{\log^{2K}(t^{\theta})\log^K(t)}{t^{\theta}}$ is bounded, and the function $\frac{\log^{2K}(f_0(\delta)/\delta)}{f_0(\delta)}$ is also uniformly bounded since it is continuous and takes 0 and $1/e$ when δ approaching to 0 and 1, respectively. We can then establish the proposition by taking a sufficiently large C_0 to guarantee that

$$\frac{\log^{2K}(C_0)}{C_0}\frac{e^{K+1}}{K^K}\frac{\log^{2K}(f_0(\delta)/\delta)}{f_0(\delta)}\sum_{t=1}^{\infty}\frac{\log^{2K}(t^{\theta})\log^K(t)}{t^{\theta}}\leq 1.$$

■

E.3 Proof of Theorem 2

By Proposition 7, those arms not in $\text{supp}(\mathcal{I})$ are pulled with sub-linear times with respect to arms in $P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})$) which will only influence $o(\log(1/\delta))$ cumulative cost and will not affect the result. Thus, in the proof of the upper bound theorem, we can assume $\text{supp}(\mathcal{I}(\boldsymbol{\mu}))=\mathcal{A}$. So, $P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})=P(\boldsymbol{c})$, $N_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})=N(\boldsymbol{c})$ and $\mathbb{E}^{\mathcal{G}}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_{\delta})]=\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_{\delta})]$

We first introduce the concept of the sample distribution function, we say

$$g : \mathbb{R}^K \times \mathbb{R}^K \to \Sigma_K \tag{16}$$

is a sample distribution, if $g(\hat{c}(t), \hat{\mu}(t))$ converge to some value g^* in Σ_K when $t \to \infty$.

For any sample distribution function $g : \mathbb{R}^K \times \mathbb{R}^K \to \Sigma_K$, $g(\hat{c}(t), \hat{\mu}(t))$ calculates the desired ratio of pulling arms, and the pulling strategy will be determined by the desired pulling ratio. As $t \to \infty$, we have $\hat{c}(t) \to c$, $\hat{\mu} \to \mu$ and $g(\hat{c}(t), \hat{\mu}(t))$ converges to the desired pulling ratios g^*.

Specifically, the sampling rule under the specific sample distribution function is as follows: we still calculate the L^∞ projection of g denoted by g^ϵ. The arm pulled at time step $t+1$ is $A_{t+1} \in \underset{1 \le a \le K}{\arg\max} \sum_{s=0}^{t} g_a^{\epsilon_s}(\hat{c}(s), \hat{\mu}(s)) - N_a(t)$. The projection allows the sampling rule to have enough exploration power by ϵ. In this case, $\sum_{s=0}^{t} g_a^{\epsilon_s}(\hat{c}(s), \hat{\mu}(s))$ is the desired number of pulls of arm a and $N_a(t)$ is the actual number of pulls of arm a. The sampling strategy

$$A_{t+1} \in \underset{1 \le a \le K}{\arg\max} \sum_{s=0}^{t} g_a^{\epsilon_s}(\hat{c}(s), \hat{\mu}(s)) - N_a(t)$$

means to pull the arm whose amount of pulls is the most lagging behind the desired number of pulls.

Before proving the main theorem, we need to have some preparations. First, we consider using the sampling rule under the sample distribution function g discussed in Equation (16). As for the lower bound, repeat the discussion in Section 4, if g converges to $\boldsymbol{\omega} \in \Sigma_K$ we have

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \ge T^*(\boldsymbol{\omega}, \boldsymbol{\mu}) \cdot \mathrm{kl}(\delta, 1-\delta)$$

where

$$T^*(\boldsymbol{\omega}, \boldsymbol{\mu})^{-1} = \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} \omega_a d(\mu_a, \lambda_a) \right) = \min_{(a,b) \in \mathcal{I}} (\omega_b + \omega_a) I_{\frac{\omega_b}{\omega_b + \omega_a}}(\mu_b, \mu_a).$$

Remark 1 *We say* $\mathrm{Alt}(\boldsymbol{\mu})$ *with out* $\boldsymbol{c}$ *means that* $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in S : \ \varphi(\boldsymbol{\lambda}) \neq \varphi(\boldsymbol{\mu})\}$.

Theorem 3 (upper bound for sample complexity) *Let* $\boldsymbol{\mu}$ *be an exponential family bandit model. Let* $\theta \in [1, e/2]$ *and* $r(t) = O(t^\theta)$. *Using stopping rule given in Equation* (11) *with* $\beta(t, \delta) = \log(r(t)/\delta)$, *and the sampling rule under any sample distribution function* g, *we have*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \le \theta T^*(g^*, \boldsymbol{\mu}).$$

This theorem is proved in Theorem 14 of Garivier and Kaufmann (2016), with some notation changes and model assumption change as $\xi = \xi(\epsilon) < \min_{(a,b) \in \mathcal{I}} |\mu_a - \mu_b|/2$, $\mathcal{I}_\epsilon := [\mu_1 - \xi, \mu_1 + \xi] \times \cdots \times [\mu_K - \xi, \mu_K + \xi]$, and the stopping statistic in the stopping rule given in Equation (11) changes to

$$\begin{aligned}
\min_{(a,b) \in \mathcal{I}} |Z_{a,b}(t)| &= \min[\min_{\substack{\mu_a < \mu_b \\ (a,b) \in \mathcal{I}}} Z_{b,a}, \min_{\substack{\mu_a > \mu_b \\ (a,b) \in \mathcal{I}}} Z_{a,b}(t)] \\
&= \min \Big[\min_{\substack{\mu_a < \mu_b \\ (a,b) \in \mathcal{I}}} \big[N_b(t) d(\hat{\mu}_b(t), \hat{\mu}_{b,a}(t)) + N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{b,a}(t)) \big], \\
&\qquad \min_{\substack{\mu_a > \mu_b \\ (a,b) \in \mathcal{I}}} \big[N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{a,b}(t)) + N_b(t) d(\hat{\mu}_b(t), \hat{\mu}_{a,b}(t)) \big] \Big].
\end{aligned}$$

Also, the function is rewritten as

$$\hat{g}(\boldsymbol{\mu}', \boldsymbol{\omega}') = \min_{(a,b) \in \mathcal{I}} (\omega_b' + \omega_a') I_{\frac{\omega_b'}{\omega_b' + \omega_a'}}(\mu_b', \mu_a')$$

using $(\omega'_b+\omega'_a)I_{\frac{\omega'_b}{\omega'_b+\omega'_a}}(\mu'_b,\mu'_a)=(\omega'_a+\omega'_b)I_{\frac{\omega'_a}{\omega'_a+\omega'_b}}(\mu'_a,\mu'_b)$.

Now, we give the proof of the Lemma 1 of the main paper:

$$\lim_{\delta\to 0} T^*(\boldsymbol{u}^*_\alpha,\boldsymbol{\mu})^{-1}\frac{\mathbb{E}_\mu[\tau_\delta]}{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]}=T^*(\boldsymbol{c},\boldsymbol{\mu})^{-1}. \tag{17}$$

Since those arms not in supp($\mathcal{I}(\boldsymbol{\mu})$) incur only sub-linear regret (with respect to arms in $P_{\mathcal{G}}(\boldsymbol{c},\boldsymbol{\mu})$), which will only influence $o(\log(1/\delta))$ of the cumulative cost, we can assume supp$(\mathcal{I}(\boldsymbol{\mu}))=\mathcal{A}$ in this proof.

Proof: Recall $\boldsymbol{u}^*_\alpha=(u_1,\dots,u_K)$ and $\boldsymbol{u}^*(\boldsymbol{c},\boldsymbol{\mu})=(u^*_1,\dots,u^*_K)$. When $\delta\to 0$, the pulling time $N_a(\tau_\delta)\to\infty$, so the sample distribution converges to $\boldsymbol{u}_\alpha$. Thus, we get

$$\lim_{\delta\to 0}\frac{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]}{\mathbb{E}_\mu[\tau_\delta]}=\frac{\sum_{i\in P(\boldsymbol{c})}c_i\mathbb{E}[N_i(\tau_\delta)]}{\sum_{i=1}^K\mathbb{E}[N_i(\tau_\delta)]}=(1-\alpha)(\sum_{i\in P(\boldsymbol{c})}u^*_ic_i).$$

Also, when $\delta\to 0$, α will converge to 1. For each $t\in P(\boldsymbol{c}),\ s\in N(\boldsymbol{c})$, by the continuity of $d(x,y)$, we have

$$\begin{aligned}
&\lim_{\alpha\to 1}\frac{\alpha/|N(\boldsymbol{c})|+u_t}{1-\alpha}I_{\frac{\alpha/|N(\boldsymbol{c})|}{\alpha/|N(\boldsymbol{c})|+u_t}}(\mu_s,\mu_t)=\lim_{\alpha\to 1}\frac{1}{(1-a)|N(\boldsymbol{c})|}I_{\frac{\alpha}{\alpha+u_t|N(\boldsymbol{c})|}}(\mu_s,\mu_t)\\
=&\lim_{\alpha\to 1}\left(\frac{1}{(1-\alpha)|N(\boldsymbol{c})|}d(\mu_s,\frac{\alpha}{\alpha+u_t|N(\boldsymbol{c})|}\mu_s+\frac{u_t|N(\boldsymbol{c})|}{\alpha+u_t|N(\boldsymbol{c})|}\mu_t)+\right.\\
&\left.\frac{1}{(1-\alpha)|N(\boldsymbol{c})|}\frac{u_t|N(\boldsymbol{c})|}{\alpha+u_t|N(\boldsymbol{c})|}d(\mu_t,\frac{\alpha}{\alpha+u_t|N(\boldsymbol{c})|}\mu_s+\frac{u_t|N(\boldsymbol{c})|}{\alpha+u_t|N(\boldsymbol{c})|}\mu_t)\right)\\
=&\lim_{\alpha\to 1}\frac{1}{(1-\alpha)|N(\boldsymbol{c})|}d(\mu_s,\frac{\alpha}{\alpha+u_t|N(\boldsymbol{c})|}\mu_s+\frac{u_t|N(\boldsymbol{c})|}{\alpha+u_t|N(\boldsymbol{c})|}\mu_t)+\frac{u_t}{(1-\alpha)}d(\mu_t,\mu_s)\\
=&\lim_{\alpha\to 1}\frac{1}{(1-\alpha)|N(\boldsymbol{c})|}d(\mu_s,\frac{\alpha}{\alpha+u_t|N(\boldsymbol{c})|}\mu_s+\frac{u_t|N(\boldsymbol{c})|}{\alpha+u_t|N(\boldsymbol{c})|}\mu_t)+u^*_td(\mu_t,\mu_s)
\end{aligned}$$

Using L'Hopital rule and $\frac{d}{dy}d(x,y)=(y-x)/b''(b^{-1}(y))$, the first term becomes

$$\lim_{\alpha\to 1}-\frac{(\mu_t-\mu_s)^2u_t^2|N(\boldsymbol{c})|^2}{(a+u_tN(\boldsymbol{c}))^3}\cdot\frac{1}{b''(b^{-1}(y))}/(-1)=0.$$

Since $u_t=(1-\alpha)u^*_t\to 0$, we have

$$\lim_{\alpha\to 1}\frac{\alpha/|N(\boldsymbol{c})|+u_t}{1-\alpha}I_{\frac{\alpha/|N(\boldsymbol{c})|}{\alpha/|N(\boldsymbol{c})|+u_t}}(\mu_s,\mu_t)=u^*_td(\mu_t,\mu_s).$$

Respectively, we have

$$\lim_{\alpha\to 1}\frac{u_s+\alpha/|N(\boldsymbol{c})|}{1-\alpha}I_{\frac{u_s}{u_s+\alpha/|N(\boldsymbol{c})|}}(\mu_s,\mu_t)=u^*_sd(\mu_s,\mu_t).$$

Now, we can compute the concerning formula as

$$\begin{aligned}
&\lim_{\delta\to 0}T^*(\boldsymbol{u}^*_\alpha,\boldsymbol{\mu})^{-1}\frac{\mathbb{E}_\mu[\tau_\delta]}{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]}\\
=&T^*(\boldsymbol{u}^*_\alpha,\boldsymbol{\mu})^{-1}\left[(1-\alpha)(\sum_{i\in P(\boldsymbol{c})}u^*_ic_i)\right]^{-1}\\
=&\lim_{\alpha\to 1}\min\left[\,T_{11}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1},T_{12}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1},T_{21}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1}\right]\cdot\left[(1-\alpha)(\sum_{i\in P(\boldsymbol{c})}u^*_ic_i)\right]^{-1}
\end{aligned}$$

where

$$T_{11}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1}=\min_{\substack{(a,b)\in\mathcal{I}\\ a,b\notin N(\boldsymbol{c})}}(u_b+u_a)I_{\frac{u_b}{u_b+u_a}}(\mu_b,\mu_a),$$

$$
\begin{aligned}
T_{12}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1} &= \min_{\substack{(a,b)\in\mathcal{I}\\ a\notin N(\boldsymbol{c}),\ b\in N(\boldsymbol{c})}} \left(\frac{\alpha}{|N(\boldsymbol{c})|}+u_a\right)I_{\frac{\alpha/|N(\boldsymbol{c})|}{\alpha/|N(\boldsymbol{c})|+u_a}}(\mu_b,\mu_a),\\
T_{21}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1} &= \min_{\substack{(a,b)\in\mathcal{I}\\ b\notin N(\boldsymbol{c}),\ a\in N(\boldsymbol{c})}} \left(u_b+\frac{\alpha}{|N(\boldsymbol{c})|}\right)I_{\frac{u_b}{u_b+\alpha/|N(\boldsymbol{c})|}}(\mu_b,\mu_a).
\end{aligned}
$$

From the previous discussion, we have

$$
\begin{aligned}
&\lim_{\alpha\to 1} T_{11}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1}\cdot\left[(1-\alpha)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)\right]^{-1}\\
&\qquad= \min_{\substack{(a,b)\in\mathcal{I}\\ a,b\notin N(\boldsymbol{c})}} (u^*_b+u^*_a)I_{\frac{u^*_b}{u^*_b+u^*_a}}(\mu_b,\mu_a)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)^{-1}\\
&\lim_{\alpha\to 1} T_{12}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1}\cdot\left[(1-\alpha)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)\right]^{-1}\\
&\qquad= \lim_{\alpha\to 1}\min_{\substack{(a,b)\in\mathcal{I}\\ a\notin N(\boldsymbol{c}),\ b\in N(\boldsymbol{c})}} \frac{\alpha/|N(\boldsymbol{c})|+u_a}{1-\alpha}I_{\frac{\alpha/|N(\boldsymbol{c})|}{\alpha/|N(\boldsymbol{c})|+u_a}}(\mu_b,\mu_a)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)^{-1}\\
&\qquad= \min_{\substack{(a,b)\in\mathcal{I}\\ a\notin N(\boldsymbol{c}),\ b\in N(\boldsymbol{c})}} u^*_a d(\mu_a,\mu_b)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)^{-1}\\
&\lim_{\alpha\to 1} T_{21}(\boldsymbol{c},\boldsymbol{u}^*_\alpha)^{-1}\cdot\left[(1-\alpha)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)\right]^{-1}\\
&\qquad= \lim_{\alpha\to 1}\min_{\substack{(a,b)\in\mathcal{I}\\ b\notin N(\boldsymbol{c}),\ a\in N(\boldsymbol{c})}} \frac{u_b+\alpha/|N(\boldsymbol{c})|}{1-\alpha}I_{\frac{u_b}{u_b+\alpha/|N(\boldsymbol{c})|}}(\mu_b,\mu_a)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)^{-1}\\
&\qquad= \min_{\substack{(a,b)\in\mathcal{I}\\ b\notin N(\boldsymbol{c}),\ a\in N(\boldsymbol{c})}} u^*_b d(\mu_b,\mu_a)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)^{-1}
\end{aligned}
$$

and combining

$$
\begin{aligned}
\left(\frac{\omega^*_b}{c_b}+\frac{\omega^*_a}{c_a}\right)\Bigg(\sum u^*_i c_i\Bigg) &= \left(\frac{\omega^*_b}{c_b}+\frac{\omega^*_a}{c_a}\right)\cdot\frac{1}{\sum_{i\in P(\boldsymbol{c})}\frac{\omega^*_i}{c_i}} = u^*_a+u^*_b,\ a,b\in P(\boldsymbol{c})\\
\frac{\omega^*_a}{c_a}\Bigg(\sum u^*_i c_i\Bigg) &= \frac{\omega^*_a}{c_a}\cdot\frac{1}{\sum_{i\in P(\boldsymbol{c})}\frac{\omega^*_i}{c_i}} = u^*_a,\ a\in P(\boldsymbol{c})
\end{aligned}
$$

we get

$$
\lim_{\delta\to 0} T^*(\boldsymbol{u}^*_\alpha,\boldsymbol{\mu})^{-1}\left[(1-\alpha)\Bigg(\sum_{i\in P(\boldsymbol{c})} u^*_i c_i\Bigg)\right]^{-1} = T^*(\boldsymbol{c},\boldsymbol{\mu})^{-1}
$$

which can be written as

$$
\lim_{\delta\to 0} T^*(\boldsymbol{u}^*_\alpha,\boldsymbol{\mu})^{-1}\frac{\mathbb{E}_\mu[\tau_\delta]}{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]} = T^*(\boldsymbol{c},\boldsymbol{\mu})^{-1}.
$$

■

After proving this lemma, combining Theorem 3, we can finish the proof of the upper bound theorem (which is Theorem 2 in the main paper).

Proof: We directly compute the result:

$$
\limsup_{\delta\to 0}\frac{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]}{\log(1/\delta)} = \limsup_{\delta\to 0}\frac{\mathbb{E}_\mu[\tau_\delta]}{\log(1/\delta)}\cdot\frac{\mathbb{E}[f(\boldsymbol{c},\boldsymbol{\mu};\tau_\delta)]}{\mathbb{E}_\mu[\tau_\delta]}
$$

$$\begin{aligned}&\le \lim_{\delta\to 0} \theta T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu}) \frac{\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)]}{\mathbb{E}_\mu[\tau_\delta]} \\ &= \theta T^*(\boldsymbol{c}, \boldsymbol{\mu})\end{aligned}$$

which concludes the proof. ■

F ALMOST-SURE UPPER BOUND

In this section, we derive the almost-sure upper bound which, as Theorem 2 indicates, also characterizes the properties of cumulative cost.

Theorem 4 (Almost Sure Upper Bound) *Let $\theta \in [1, e/2]$ and $r(t) = O(t^\theta)$. Using stopping rule given in Equation* (11) *with $\beta(t, \delta) = \log(r(t)/\delta)$, the CAET algorithm ensures*

$$\mathbb{P}_\mu \left(\limsup_{\delta\to 0} \frac{f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)}{\log(1/\delta)} \le \theta T^*(\boldsymbol{c}, \boldsymbol{\mu}) \right) = 1.$$

Proof: For simplicity, we still assume $\mathrm{supp}(\mathcal{I}(\boldsymbol{\mu})) = \mathcal{A}$ with the reason stated as before. When t goes to infinity, the actual sampling distribution is

$$\boldsymbol{u}^*_\alpha = \boldsymbol{u}^*_{\alpha(\delta)} = 1_{\alpha(\delta)}(\boldsymbol{c}) + (1 - \alpha(\delta))\boldsymbol{u}^*(\boldsymbol{c}, \boldsymbol{\mu}) := (u_1, u_2, \ldots, u_K).$$

Define $\mathcal{E}$ as

$$\mathcal{E} = \{\forall a \in \mathcal{A}, N_a(t)/t \underset{t\to\infty}{\to} u_a\} \cup \{\hat{\boldsymbol{\mu}}(t) \underset{t\to\infty}{\to} \boldsymbol{\mu}\} \cup \{\hat{\boldsymbol{c}}(t) \underset{t\to\infty}{\to} \boldsymbol{c}\}.$$

From the Proposition 3 and the Law of Large Numbers, event $\mathcal{E}$ happens with probability 1. We have:

$$\frac{f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)}{\tau_\delta} = \sum_{a\in P(\boldsymbol{c})} c_a \frac{N_a(\tau_\delta)}{\tau_\delta}.$$

And, in event $\mathcal{E}$, $N_a(t)/t \underset{t\to\infty}{\to} u_a$ and τ_δ will goes to ∞ as the lower bound of τ_δ implies. So,

$$\mathbb{P} \left(\frac{f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)}{\tau_\delta} \xrightarrow[\delta\to 0]{} (1 - \alpha) \sum_{a\in P(\boldsymbol{c})} c_a u^*_a \right) = 1. \tag{18}$$

Define

$$T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu})^{-1} = \min_{(a,b)\in\mathcal{I}} \inf_{\substack{\lambda\in\mathcal{S}: \\ (\lambda_a-\lambda_b)(\mu_a-\mu_b)<0}} \sum_{a=1}^K u_a d(\mu_a, \lambda_a) = \min_{(a,b)\in\mathcal{I}} (u_a + u_b) I_{\frac{u_b}{u_b+u_a}}(\mu_b, \mu_a).$$

According to Proposition 13 in Garivier and Kaufmann (2016), we have τ_δ is bounded for any $\delta > 0$ and

$$\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \le (1 + \epsilon)\theta T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu}).$$

Also, Lemma 1 in the main paper, which is Equation (17) in this Appendix, tells us that

$$\lim_{\delta\to 0} T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu})^{-1}(1 - \alpha) \sum_{a\in P(\boldsymbol{c})} c_a u^*_a = T^*(\boldsymbol{c}, \boldsymbol{\mu})^{-1}.$$

Combining with Equation (18), we get

$$\begin{aligned}\limsup_{\delta\to 0} \frac{f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)}{\log(1/\delta)} &= \limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \cdot \frac{f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)}{\tau_\delta} \\ &\le (1 + \epsilon)\theta \limsup_{\delta\to 0} T^*(\boldsymbol{u}^*_\alpha, \boldsymbol{\mu}) \cdot \frac{f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)}{\tau_\delta} \\ &= (1 + \epsilon)\theta T^*(\boldsymbol{c}, \boldsymbol{\mu}).\end{aligned}$$

Letting ϵ go to zero concludes the proof. ■

G SPECIAL CASES

In this section, we leverage the theory developed above to study three important special cases of interest – best arm identification, ranking identification, and regret minimization, by specifying the parameters (identification task $\mathcal{I}$ and cost vector $\boldsymbol{c}$) in the general scenario.

The cost-aware scenario of BAI was studied in a concurrent work (Kanarios et al., 2024), yet for a strictly positive cost vector. In contrast, we consider more general identification problems with non-negative cost vectors. Note that allowing certain costs to be zero leads to more technical challenges and allows the generality of the scenarios. For example, the regret minimization case studied in Section 7 of the main paper, can be viewed as a case with a non-negative cost vector.

G.1 Best Arm Identification

In the best arm identification (BAI) problem, the goal is to identify the arm with the largest expected reward. In the pure-exploration problem $(\mathcal{G}, \varphi)$ of identifying the best arm, $M = K$ and $\mathcal{G}_k = \cap_{j \neq k} \mathcal{B}_{kj}$, which is a pairwise exploration task. With a positive cost vector $\boldsymbol{c} = (c_1, \dots, c_K)$, our theorem implies, for any δ-PAC algorithm

$$\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)] \geq T^*(\boldsymbol{c}, \boldsymbol{\mu}) \mathrm{kl}(\delta, 1-\delta)$$

where

$$T^*(\boldsymbol{c}, \boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega} \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in \mathcal{A}} \frac{\omega_a}{c_a} d(\mu_a, \lambda_a).$$

Here $c_i > 0$, so actually $\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu}) = \mathrm{Alt}(\boldsymbol{\mu})$. And our CAET algorithm satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[f(\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta)]}{\log(1/\delta)} \leq \theta T^*(\boldsymbol{c}, \boldsymbol{\mu})$$

recovering the results in Kanarios et al. (2024). Also, when we take $\boldsymbol{c} = (1, \dots, 1)$, the cost $\mathbb{E}[\boldsymbol{c}, \boldsymbol{\mu}; \tau_\delta]$ become sample time $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ and gets

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \theta T^*(\boldsymbol{\mu}).$$

which is the result is Garivier and Kaufmann (2016). Thus, our algorithm does improve from the previous ones and can handle more situations.

G.2 Ranking Identification

Ranking identification is another important task. Instead of just identifying the best arm, the agent needs to identify the total order of all the K arms in the bandits model $\boldsymbol{\mu}$. In our theory, considering the pure-exploration problem $(\mathcal{G}, \varphi)$ of identifying the rank of all the arms with respect to their expected rewards, $M = K!$ and each $\mathcal{G}_m \in \mathcal{G}$ is a singleton containing some σ. It is a pairwise exploration task since $\{\sigma\} = \cap_{i=1}^{K-1} \mathcal{B}_{\sigma^{-1}(i)\sigma^{-1}(i+1)}$. By taking different values of cost vector $\boldsymbol{c}$, our theorem can be applied to minimize sample complexity setting and regret minimization setting, both of which are intriguing and fresh settings.

The auxiliary optimal value has an explicit expression in the ranking identification problem as illustrated in Lemma 13 below which can help us to calculate the value of $T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})$ in the theorem.

Lemma 13 *Assume $\mu_1 > \mu_2 > \dots > \mu_K$. For ranking identification task,* $\mathrm{supp}(\mathcal{I}_m) = \mathcal{A}$*, so $P_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu}) = P(\boldsymbol{c})$ and $N_{\mathcal{G}}(\boldsymbol{c}, \boldsymbol{\mu}) = N(\boldsymbol{c})$. For every $\omega \in \Sigma_{P(\boldsymbol{c})}$, We have*

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu})} \Big(\sum_{a \in P(\boldsymbol{c})} \omega_a \frac{d(\mu_a, \lambda_a)}{c_a} \Big) = \min \Bigg[\min_{a > b,\ a, b \notin N(\boldsymbol{c})} \Big(\frac{\omega_b}{c_b} + \frac{\omega_a}{c_a} \Big) I_{\frac{\omega_b / c_b}{\omega_b / c_b + \omega_a / c_a}} (\mu_b, \mu_a),$$

$$\min_{a > b,\ a \notin N(\boldsymbol{c}),\ b \in N(\boldsymbol{c})} \Big(\frac{\omega_a}{c_a} d(\mu_a, \mu_b) \Big), \min_{a > b,\ b \notin N(\boldsymbol{c}),\ a \in N(\boldsymbol{c})} \Big(\frac{\omega_b}{c_b} d(\mu_b, \mu_a) \Big) \Bigg].$$

Proof: Let $\boldsymbol{\mu}$ such that $\mu_1 > \mu_2 > \ldots > \mu_K$ and we can note that $\mathrm{Alt}(\boldsymbol{c}, \boldsymbol{\mu}) = \underset{a>b}{\cup}\{\boldsymbol{\lambda} \in S \ : \ \lambda_a > \lambda_b, \ \lambda_i = \mu_i \ \forall i \in N(\boldsymbol{c})\}$. Using that fact, one has

$$\begin{aligned}
T^*(\boldsymbol{c},\boldsymbol{\mu})^{-1} &= \sup_{\omega\in\Sigma_{P(\boldsymbol{c})}} \min_{a>b} \inf_{\boldsymbol{\lambda}\in S:\lambda_a>\lambda_b, \ \lambda_i=\mu_i \ \forall i\in N(\boldsymbol{c})} \left(\sum_{a\in P(\boldsymbol{c})} \frac{\omega_a}{c_a} d(\mu_a,\lambda_a)\right) \\
&= \sup_{\omega\in\Sigma_{P(\boldsymbol{c})}} \min_{a>b} \inf_{\boldsymbol{\lambda}\in S:\lambda_a\geq\lambda_b, \ \lambda_i=\mu_i \ \forall i\in N(\boldsymbol{c})} \left(\sum_{a\in P(\boldsymbol{c})} \frac{\omega_a}{c_a} d(\mu_a,\lambda_a)\right) \\
&= \sup_{\omega\in\Sigma_{P(\boldsymbol{c})}} \min \Bigg[\min_{a>b, \ a,b\notin N(\boldsymbol{c})} \inf_{\boldsymbol{\lambda}\in S:\lambda_a\geq\lambda_b} \left(\frac{\omega_b}{c_b} d(\mu_b,\lambda_b) + \frac{\omega_a}{c_a} d(\mu_a,\lambda_a)\right), \\
&\qquad \min_{a>b, \ a\notin N(\boldsymbol{c}), \ b\in N(\boldsymbol{c})} \inf_{\boldsymbol{\lambda}\in S:\lambda_a\geq\lambda_b=\mu_b} \left(\frac{\omega_a}{c_a} d(\mu_a,\lambda_a)\right), \\
&\qquad \min_{a>b, \ b\notin N(\boldsymbol{c}), \ a\in N(\boldsymbol{c})} \inf_{\boldsymbol{\lambda}\in S:\mu_a=\lambda_a\geq\lambda_b} \left(\frac{\omega_b}{c_b} d(\mu_b,\lambda_b)\right)\Bigg] \\
&= \sup_{\omega\in\Sigma_{P(\boldsymbol{c})}} \min \Bigg[\min_{a>b, \ a,b\notin N(\boldsymbol{c})} \inf_{\boldsymbol{\lambda}\in S:\lambda_a\geq\lambda_b} \left(\frac{\omega_b}{c_b} d(\mu_b,\lambda_b) + \frac{\omega_a}{c_a} d(\mu_a,\lambda_a)\right), \\
&\qquad \min_{a>b, \ a\notin N(\boldsymbol{c}), \ b\in N(\boldsymbol{c})} \left(\frac{\omega_a}{c_a} d(\mu_a,\mu_b)\right), \min_{a>b, \ b\notin N(\boldsymbol{c}), \ a\in N(\boldsymbol{c})} \left(\frac{\omega_b}{c_b} d(\mu_b,\mu_a)\right)\Bigg]
\end{aligned}$$

Minimizing

$$f(\lambda_b,\lambda_a) = \frac{\omega_b}{c_b} d(\mu_b,\lambda_b) + \frac{\omega_a}{c_a} d(\mu_a,\lambda_a)$$

under the constraint $\lambda_a \geq \lambda_b$ is a convex optimization problem that can be solved analytically. The minimum is obtained for

$$\lambda_b = \lambda_a = \frac{\omega_b/c_b}{\omega_b/c_b + \omega_a/c_a}\mu_b + \frac{\omega_a/c_a}{\omega_b/c_b + \omega_a/c_a}\mu_a$$

and its value can be written $(\frac{\omega_b}{c_b} + \frac{\omega_a}{c_a}) I_{\frac{\omega_b/c_b}{\omega_b/c_b+\omega_a/c_a}}(\mu_b,\mu_a)$.

So we have

$$\begin{aligned}
T^*(\boldsymbol{c},\boldsymbol{\mu})^{-1} = \sup_{\omega\in\Sigma_{P(\boldsymbol{c})}} \min \Bigg[&\min_{a>b,a,b\notin N(\boldsymbol{c})} \left(\frac{\omega_b}{c_b} + \frac{\omega_a}{c_a}\right) I_{\frac{\omega_b/c_b}{\omega_b/c_b+\omega_a/c_a}}(\mu_b,\mu_a), \\
&\min_{a>b,a\notin N(\boldsymbol{c}), \ b\in N(\boldsymbol{c})} \left(\frac{\omega_a}{c_a} d(\mu_a,\mu_b)\right), \min_{a>b,b\notin N(\boldsymbol{c}), \ a\in N(\boldsymbol{c})} \left(\frac{\omega_b}{c_b} d(\mu_b,\mu_a)\right)\Bigg].
\end{aligned}$$

■

For some closed-form expressions, we give an example of ranking identification under regret minimization in three-armed bandit, we can present its closed-form clearly. For a three-armed 1-Gaussian bandit model $\boldsymbol{\mu} = (\mu_1,\mu_2,\mu_3)$, $\boldsymbol{c} = \boldsymbol{\Delta} = (0,\Delta_2,\Delta_3)$, $T^*(\boldsymbol{\Delta},\boldsymbol{\mu})$ and $\boldsymbol{u}^*(\boldsymbol{\Delta},\boldsymbol{\mu}) = G_{\boldsymbol{\Delta}}(\boldsymbol{\omega}^*(\boldsymbol{\Delta},\boldsymbol{\mu}))$ are shown in next proposition.

Proposition 8 *Let $\delta \in (0,1)$. For any δ-PAC algorithm under ranking identification with cost vector $\boldsymbol{\Delta}$. Any 3-armed 1-Gaussian bandits with reward expectations $\{\mu_1 > \mu_2 > \mu_3\}$ has:*

$$\boldsymbol{u}^*(\boldsymbol{\Delta},\boldsymbol{\mu}) = \begin{cases} (0, \frac{\sqrt{\Delta_3}}{\sqrt{\Delta_2}+\sqrt{\Delta_3}}, \frac{\sqrt{\Delta_2}}{\sqrt{\Delta_2}+\sqrt{\Delta_3}}) & \Delta_3 \leq [(3+\sqrt{5})/2]\Delta_2 \\ (0, \frac{\Delta_3^2-2\Delta_2\Delta_3}{(\Delta_3-\Delta_2)^2}, \frac{\Delta_2^2}{(\Delta_3-\Delta_2)^2}) & \Delta_3 > [(3+\sqrt{5})/2]\Delta_2 \end{cases}$$

$$T^*(\boldsymbol{\Delta},\boldsymbol{\mu})^{-1} = \begin{cases} \frac{(\sqrt{\Delta_2}-\sqrt{\Delta_3})^2}{2} & \Delta_3 \leq [(3+\sqrt{5})/2]\Delta_2 \\ \frac{\Delta_2}{2}\frac{\Delta_3-2\Delta_2}{\Delta_3-\Delta_2} & \Delta_3 > [(3+\sqrt{5})/2]\Delta_2 \end{cases}$$

and $R(\tau_\delta) \geq T^*(\boldsymbol{\Delta},\boldsymbol{\mu})\mathrm{kl}(\delta, 1-\delta)$.

Proof: First, consider the expression of $T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})^{-1}$ and the setting is $\mu_1 > \mu_2 > \mu_3,\ \Delta_i = \mu_1 - \mu_i$ is the sub-optimal gap. Recall that $d(x, y) = (x - y)^2/2$ under 1-Gaussian bandit setting, so

$$\begin{aligned} T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})^{-1} &= \sup_{\omega_2+\omega_3=1} \min\big[\frac{\omega_2}{\Delta_2} d(\mu_2, \mu_1), \frac{\omega_3}{\Delta_3} d(\mu_3, \mu_1), (\frac{\omega_2}{\Delta_2} + \frac{\omega_3}{\Delta_3}) I_{\frac{\omega_2/\Delta_2}{\omega_2/\Delta_2+\omega_3/\Delta_3}}(\mu_2, \mu_3)\big] \\ &= \sup_{\omega_2+\omega_3=1} \min\big[\frac{\omega_2\Delta_2}{2}, \frac{\omega_3\Delta_3}{2}, (\frac{\omega_2}{\Delta_2} + \frac{\omega_3}{\Delta_3}) I_{\frac{\omega_2\Delta_3}{\omega_2\Delta_3+\omega_3\Delta_2}}(\mu_2, \mu_3)\big]. \end{aligned}$$

Noticing that

$$\begin{aligned} (\frac{\omega_2}{\Delta_2} + \frac{\omega_3}{\Delta_3}) I_{\frac{\omega_2/\Delta_2}{\omega_2/\Delta_2+\omega_3/\Delta_3}}(\mu_2, \mu_3) &= \inf_{\lambda_3 \geq \lambda_2} (\frac{\omega_2}{\Delta_2} d(\mu_2, \lambda_2) + \frac{\omega_3}{\Delta_3} d(\mu_3, \lambda_3)) \\ &\leq_{\lambda_3=\mu_1, \lambda_2=\mu_2} (\frac{\omega_2}{\Delta_2} d(\mu_2, \lambda_2) + \frac{\omega_3}{\Delta_3} d(\mu_3, \lambda_3)) \\ &= \frac{\omega_3}{\Delta_3} d(\mu_3, \mu_1), \end{aligned}$$

we do not need to take term $\frac{\omega_3}{\Delta_3} d(\mu_3, \mu_1)$ into consideration in computing $T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})^{-1}$. Let $\omega_3 = 1 - \omega_2$ and

$$(\frac{\omega_2}{\Delta_2} + \frac{\omega_3}{\Delta_3}) I_{\frac{\omega_2\Delta_3}{\omega_2\Delta_3+\omega_3\Delta_2}}(\mu_2, \mu_3) = \frac{(\mu_2 - \mu_3)^2}{2} \frac{\omega_2(1 - \omega_2)}{\Delta_3\omega_2 + \Delta_2(1 - \omega_2)}.$$

Define function

$$h(x) = \frac{(\mu_2 - \mu_3)^2}{2} \frac{x(1 - x)}{\Delta_3 x + \Delta_2(1 - x)},\ x \in [0, 1].$$

We need to discover some properties of $h(x)$

$$h'(x) = \frac{(\mu_2 - \mu_3)^2}{2} \frac{-\Delta_3 x^2 + \Delta_2(x - 1)^2}{(-\Delta_3 x + \Delta_2(x - 1))^2}$$

and

$$h''(x) = \frac{(\mu_2 - \mu_3)^2}{2} \frac{2\Delta_2\Delta_3}{(-\Delta_3 x + \Delta_2(x - 1))^3}.$$

Since $-\Delta_3 x + \Delta_2(x - 1) < 0$, so $h''(x) < 0$ which means $h(x)$ is convex in $[0, 1]$ and $h'(x) = 0$ shows that $h(x)$ achieve maximum value $h_m = \frac{(\mu_2-\mu_3)^2}{2(\sqrt{\Delta_2}+\sqrt{\Delta_3})^2} = \frac{(\sqrt{\Delta_2}-\sqrt{\Delta_3})^2}{2}$ at point $x = \frac{\sqrt{\Delta_2}}{\sqrt{\Delta_2}+\sqrt{\Delta_3}}$. After establishing those properties, we can continue our discussion. When $h_m \leq \frac{\Delta_2}{2} \frac{\sqrt{\Delta_2}}{\sqrt{\Delta_2}+\sqrt{\Delta_3}}$ which means $(\mu_2 - \mu_3)^2 \leq \Delta_2^{3/2}(\sqrt{\Delta_2} + \sqrt{\Delta_3})$. Solving this inequality with condition $\Delta_3 > \Delta_2$, we get $\Delta_3 \leq [(3 + \sqrt{5})/2]\Delta_2$, and further have:

$$T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})^{-1} = h_m = \frac{(\sqrt{\Delta_2} - \sqrt{\Delta_3})^2}{2}.$$

When $h_m > \frac{\Delta_2}{2} \frac{\sqrt{\Delta_2}}{\sqrt{\Delta_2}+\sqrt{\Delta_3}}$, which means $\Delta_3 > [(3 + \sqrt{5})/2]\Delta_2$, let x_0 be the solution of $h(x) = \frac{\Delta_2}{2} x$ and we get:

$$T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})^{-1} = \frac{\Delta_2}{2} x_0$$

where

$$x_0 = \frac{(\mu_2 - \mu_3)^2 - \Delta_2^2}{(\mu_2 - \mu_3)^2 + \Delta_2\Delta_3 - \Delta_2^2} = \frac{\Delta_3 - 2\Delta_2}{\Delta_3 - \Delta_2}.$$

■

G.3 Proof of Equation (8) in the Main Paper

Assume μ_1 is the unique best arm, with confidence δ and $\varphi(\boldsymbol{\mu}) = 1, \mathcal{I}_1 = \{(1,a) : a = 2, \ldots, K\}$, and $\operatorname{supp}(\mathcal{I}(\boldsymbol{\mu})) = \mathcal{A}$. Our theorem implies that

$$R(\tau_\delta) \leq \theta T^*(\boldsymbol{\Delta}, \boldsymbol{\mu}) \log\left(\frac{1}{\delta}\right)$$

where $\boldsymbol{\Delta} = (0, \Delta_1, \ldots, \Delta_K)$ and using Lemma 3 in Appendix C.2

$$T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})^{-1} = \sup_{\substack{\omega_2 + \cdots + \omega_K = 1 \\ \omega_i > 0}} \min_{a \in \{2, \ldots, K\}} \left\{ \frac{\omega_a}{\Delta_a} d(\mu_a, \mu_1) \right\}.$$

To solve the final optimization problem, we use some inequalities:

$$\begin{aligned} \min_{a \in \{2, \ldots, K\}} \left\{ \frac{\omega_a}{\Delta_a} d(\mu_a, \mu_1) \right\} &\leq \sum_{a=2}^{K} \frac{\omega_a}{\Delta_a} d(\mu_a, \mu_1) \cdot \frac{\frac{\Delta_a}{d(\mu_a, \mu_1)}}{\sum\limits_{i=2}^{K} \frac{\Delta_i}{d(\mu_i, \mu_1)}} \\ &= \sum_{a=2}^{K} \omega_a \cdot \frac{1}{\sum\limits_{i=2}^{K} \frac{\Delta_i}{d(\mu_i, \mu_1)}}. \end{aligned}$$

The first inequality comes from the fact that $\min\{a_1, \ldots, a_n\} \leq \sum_i p_i a_i$, if $\sum_i p_i = 1, p_i \geq 0$. In the last term, we have $\sum_a \omega = 1$, so

$$\min_{a \in \{2, \ldots, K\}} \left\{ \frac{\omega_a}{\Delta_a} d(\mu_a, \mu_1) \right\} \leq \frac{1}{\sum\limits_{i=2}^{K} \frac{\Delta_i}{d(\mu_i, \mu_1)}}.$$

The right-hand side is a constant and can be obtained by

$$(\omega_2, \ldots, \omega_K) = \left(\frac{\frac{\Delta_2}{d(\mu_2, \mu_1)}}{\sum\limits_{i=2}^{K} \frac{\Delta_i}{d(\mu_i, \mu_1)}}, \ldots, \frac{\frac{\Delta_K}{d(\mu_K, \mu_1)}}{\sum\limits_{i=2}^{K} \frac{\Delta_i}{d(\mu_i, \mu_1)}} \right).$$

So, we have

$$T^*(\boldsymbol{\Delta}, \boldsymbol{\mu})^{-1} = \sup_{\substack{\omega_2 + \cdots + \omega_K = 1 \\ \omega_i > 0}} \min_{a \in \{2, \ldots, K\}} \left\{ \frac{\omega_a}{\Delta_a} d(\mu_a, \mu_1) \right\} = \frac{1}{\sum\limits_{i=2}^{K} \frac{\Delta_i}{d(\mu_i, \mu_1)}},$$

and the regret in our ETC algorithm is

$$R_{\boldsymbol{\mu}}(T) \leq \underbrace{\theta T^*(\boldsymbol{\Delta}, \boldsymbol{\mu}) \log(T)}_{\text{regret before commitment}} + \underbrace{(T - O(\log^{1+r}(T))) \frac{1}{T} \max_a \Delta_a}_{\text{regret during commitment}}.$$

The last term can be bounded by a constant, and we get

$$R_{\boldsymbol{\mu}}(T) \leq \theta \sum_{i=2}^{K} \frac{\Delta_i}{d(\mu_i, \mu_1)} \log(T) = \theta \sum_{i=2}^{K} \frac{\Delta_i}{\mathrm{KL}(\mu_i, \mu_1)} \log(T)$$

where θ can be chosen to be close to 1, which concludes the proof.