Journal of Machine Learning Research 27 (2026) 1-41 Submitted 9/24; Revised 4/26; Published 5/26

Best Arm Identification with Minimal Regret

Junwen Yang JUNWEN_YANG@U.NUS.EDU
Institute of Operations Research and Analytics
National University of Singapore
117602, Singapore

Vincent Y. F. Tan VTAN@NUS.EDU.SG
Department of Mathematics
Department of Electrical and Computer Engineering
Institute of Operations Research and Analytics
National University of Singapore
119076, Singapore

Tianyuan Jin* TIANYUAN@U.NUS.EDU
Department of Mathematics
National University of Singapore
119076, Singapore

Editor: Vianney Perchet

Abstract

Motivated by real-world applications that necessitate responsible experimentation, we introduce the problem of *best arm identification (BAI) with minimal regret*. This variant of the multi-armed bandit problem elegantly amalgamates two of its most ubiquitous objectives: regret minimization and BAI. More precisely, the agent's goal is to identify the best arm with a prescribed confidence level δ, while minimizing the cumulative regret up to the stopping time. Focusing on single-parameter exponential families of distributions, we leverage information-theoretic techniques to establish an instance-dependent lower bound on the expected cumulative regret. Moreover, we present an impossibility result that underscores the tension between cumulative regret and sample complexity in fixed-confidence BAI. Complementarily, we design and analyze the Double KL-UCB algorithm, which achieves asymptotic optimality as the confidence level tends to zero. Notably, this algorithm employs two distinct confidence bounds to guide arm selection in a randomized manner. Our findings elucidate a fresh perspective on the inherent connections between regret minimization and BAI.

Keywords: multi-armed bandits, best arm identification, regret minimization

1. Introduction

The multi-armed bandit model, originally introduced by Thompson (1933), offers a straightforward yet powerful online learning framework that has undergone extensive investigation in the online decision making literature. In the domain of stochastic multi-armed bandits, the agent grapples with the intricate challenge of optimizing arm-pulling decisions to attain specific objectives. Notably, the *regret minimization* problem aims at maximizing the cumu-

*. Corresponding author.

lative rewards (or equivalently, minimizing the cumulative regret) by delicately balancing the trade-off between exploration and exploitation (Agrawal, 1995; Auer et al., 2002; Bubeck and Cesa-Bianchi, 2012). On the other hand, the *best arm identification* (BAI) problem focuses on achieving efficient exploration of the optimal arm (Even-Dar et al., 2006; Audibert et al., 2010; Karnin et al., 2013; Garivier and Kaufmann, 2016).

Nevertheless, despite being arguably two of the most prominent problems in the studies of multi-armed bandits, the performance metrics for BAI and regret minimization are distinct and merit rather different considerations in the design of algorithms and their analyses. In the context of fixed-confidence BAI, the agent seeks to identify the best arm with a given (high) confidence level. Here, the conventional emphasis lies solely on the *sample complexity*. In other words, the agent strives to utilize as few samples as possible to achieve the goal of identifying the best arm, without consideration of the cumulative regret. While widely recognized to be important, this problem formulation may not be applicable to all practical settings, as it assumes that the cost of pulling each arm is the same. In real-world scenarios, cognizant of the fact that the cost of pulling an arm is characterized by the suboptimality gap of that arm (which generally differs across arms), the agent might desire to prioritize the minimization of the *cumulative regret* during the interaction with the environment. This approach reflects a commitment to a more responsible experimental process.

Consider, for example, the field of clinical trials, where researchers aim to discover the most effective medical treatment. However, focusing solely on the quantity of trials conducted, which corresponds to sample complexity, overlooks a crucial aspect—the different impact of various treatments on patients. Consequently, the overall cumulative regret of the study emerges as a more ethically grounded and patient-centric performance metric.

In light of this disparity between existing theoretical formulations and practical considerations, we investigate the problem of *best arm identification with minimal regret*, where the expected cumulative regret up to the stopping time serves as the performance metric of BAI. Through a comprehensive theoretical examination of this crucial problem, our work not only addresses a significant gap in existing bandit studies but also offers a fresh, novel, and somewhat surprising perspective on the underlying connection between regret minimization and BAI.

Main contributions. Our main results and contributions are summarized as follows:

(i) In Section 3, we derive fundamental information-theoretic limits for the problem of BAI with minimal regret, as rigorously formulated in Section 2. Specifically, utilizing the change-of-measure technique, we establish an instance-dependent lower bound on the expected cumulative regret for all feasible online BAI algorithms. Furthermore, we derive an impossibility result, as outlined in Theorem 5. This result underscores the inevitability that achieving minimal cumulative regret in BAI necessitates a higher-order sample complexity.

(ii) In Sections 4 and 5, we propose and analyze the Double KL-UCB (or DKL-UCB) algorithm. Specifically, at each time step, DKL-UCB selects two candidate arms based on two meticulously designed Kullback–Leibler upper confidence bounds, and then employs an appropriate amount of randomization to pull one of the candidate arms.

We establish that DKL-UCB achieves asymptotic optimality, wherein its expected cumulative regret attains the lower bound as the error probability δ tends to zero. Additionally, we demonstrate that its expected sample complexity is nearly optimal, up to a small multiplicative factor that scales as the square of $\log\log(1/\delta)$.

(iii) In Section 6, we conduct a comparative analysis between our examined problem and two other classical problems: regret minimization and BAI with minimal samples, with a focus on their respective asymptotic performances. We find that the hardness parameter for regret minimization aligns seamlessly with that of our specific problem, warranting further investigation. Besides, we illustrate the significant distinctions that arise in the context of BAI when employing different performance metrics.

Related work. Both the problems of cumulative regret minimization and best arm identification have attracted considerable attention in the literature. The asymptotic lower bound for regret minimization was firstly established by Lai and Robbins (1985), and further generalized by Burnetas and Katehakis (1996). Subsequently, a diverse range of policies, such as KL-UCB (Garivier and Cappé, 2011; Cappé et al., 2013) and Thompson Sampling (Agrawal and Goyal, 2012; Korda et al., 2013), have been meticulously developed to achieve asymptotic optimality across various scenarios. For an extensive survey of regret minimization algorithms, we refer to Lattimore and Szepesvári (2020).

In this work, we focus on the fixed-confidence BAI problem (Even-Dar et al., 2002). A considerable body of research has been dedicated to establishing upper and lower bounds on the sample complexity of fixed-confidence BAI (Kalyanakrishnan et al., 2012; Gabillon et al., 2012; Jamieson et al., 2014; Kaufmann et al., 2016), and the gap was ultimately closed in Garivier and Kaufmann (2016). Specifically, this insightful work provides a complete characterization of the expected sample complexity of BAI as the confidence level goes to zero, and achieves the minimal sample complexity through a tracking strategy named Track-and-Stop.

While this paper focuses on the fixed-confidence setting, it is worth noting that BAI can also be explored under different paradigms: (1) the fixed-budget setting (Audibert et al., 2010; Karnin et al., 2013; Carpentier and Locatelli, 2016; Degenne, 2023), which consists in minimizing error probability with a predetermined exploration budget; and (2) minimization of the simple regret (Bubeck et al., 2009; Lattimore et al., 2016; Zhao et al., 2023), which consists in minimizing the expected regret of the selected arm after exploration. These frameworks are generally non-interchangeable and lead to different theoretical insights and performance guarantees.

Consistent with our intrinsic motivation, several works interpolate the objectives of regret minimization and BAI. Among these, the work most closely related to ours is Degenne et al. (2019b). In particular, a dual-objective algorithm UCB_α was introduced and analyzed, where the parameter α serves to externally balance between regret and sample complexity. However, we remark that Degenne et al. (2019b) only established achievability results, with no lower bounds provided. In contrast, our work focuses on the minimization of expected cumulative regret for BAI, and presents an algorithm that asymptotically attains the fundamental lower bound; hence, tightness and optimality of the results are ensured. Subsequently, akin to Degenne et al. (2019b), Zhong et al. (2023) quantified the trade-off between regret minimization and BAI in the fixed-budget setting. Furthermore, Zhang and

Ying (2023) studied whether the asymptotically optimal algorithms for the problem of regret minimization can commit to the best arm quickly. Their perspective is motivated by the practical preference for quick commitment over continuous exploration, and their results are not directly comparable to ours. More recently, Qin and Russo (2024) considered a model where the agent can choose to stop experimenting adaptively before reaching the total time horizon. In this model, the objective is to optimize concurrently both within-experiment and post-experiment cost functions. Kanarios et al. (2024) studied cost-aware BAI, where each arm is associated with a cost distribution with support in $[\ell, 1]$ where $\ell > 0$, and the agent aims to minimize the cumulative cost up to the stopping time. A crucial difference of our framework is that the cost associated with the optimal arm is exactly zero in the context of cumulative regret, leading to fundamentally different theoretical results.

2. Problem Setup and Preliminaries

Multi-armed bandits. We consider a conventional stochastic multi-armed bandit model with finitely many arms. Specifically, the arm set is denoted as $[K] := \{1, 2, \ldots, K\}$ and each arm i is associated with a reward distribution ν_i with mean $\mu_i \in \mathbb{R}$. We adopt the assumption that the reward distributions ν_i for all $i \in [K]$ belong to a known canonical single-parameter exponential family. A formal introduction to this family will follow shortly, with a crucial takeaway being that the bandit instance can be fully characterized by the means of its arms, expressed as the vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K) \in \mathbb{R}^K$. For convenience, we assume that arm 1 is the unique[1] optimal arm, i.e., $1 = \arg\max_{i \in [K]} \mu_i$. For each arm $i \in [K]$, we define $\Delta_i := \mu_1 - \mu_i$ as its *suboptimality gap*. Furthermore, let $\mathcal{M}$ denote the collection of all bandit instances defined above.

At each time step $t \in \mathbb{N}$, the agent chooses an arm A_t from the given arm set $[K]$. Subsequently, the agent observes a random reward X_t, which is drawn from its associated distribution ν_{A_t} and independent of the rewards obtained from previous time steps. Notably, subsequent arm selections, in general, depend on both prior arm choices and preceding rewards.

Best arm identification with minimal regret. In the fixed-confidence setting where a confidence level $\delta \in (0, 1)$ is given, the agent aims to identify the best arm with a probability of at least $1-\delta$ and minimal expected cumulative regret. This is achieved through sequential and adaptive arm pulling.

More formally, the agent employs an *online algorithm* π to decide the arm A_t to pull at each time step t, to select a time τ_δ to stop pulling arms, and to ultimately recommend i_{out} as the identified best arm to output. Let $\mathcal{F}_t := \sigma(A_1, X_1, \ldots, A_t, X_t)$ denote the σ-field generated by the interaction history up to and including time t. Thus, the online algorithm π consists of three integral components:

- The *sampling rule* selects arm A_t, which is $\mathcal{F}_{t-1}$-measurable;
- The *stopping rule* determines a stopping time τ_δ, which is adapted to the filtration $(\mathcal{F}_t)_{t=1}^{\infty}$;
- The *recommendation rule* produces a candidate best arm i_{out}, which is $\mathcal{F}_{\tau_\delta}$-measurable.

1. The uniqueness assumption is commonly applied in fixed-confidence BAI, as it is impossible to distinguish between two arms with identical means.

Definition 1 *For a prescribed confidence level $\delta \in (0,1)$, an online best arm identification algorithm π is said to be* δ-PAC (probably approximately correct) *if, for all bandit instances $\boldsymbol{\mu} \in \mathcal{M}$, it terminates within a finite time almost surely and the probability of error is no more than δ, i.e., $\mathbb{P}_{\boldsymbol{\mu}}(i_{\text{out}} \neq 1) \leq \delta$.*

Let $R(t) := \sum_{s=1}^{t}(\mu_1 - X_s)$ represent the cumulative regret up to any time step $t \in \mathbb{N}$. Then our overarching goal is to design and analyze a δ-PAC online BAI algorithm while minimizing its expected cumulative regret $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]$. Essentially, we seek to address the following problem:

$$\begin{aligned} \min \quad & \mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] \\ \text{s.t.} \quad & \mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty) = 1 \text{ and } \mathbb{P}_{\boldsymbol{\mu}}(i_{\text{out}} \neq 1) \leq \delta \end{aligned} \tag{1}$$

where $R(\tau_\delta) = \sum_{t=1}^{\tau_\delta}(\mu_1 - X_t)$ and the minimization is over all online algorithms as defined above.

Remark 2 *The problem under investigation is closely related to two classical challenges in the multi-armed bandit literature: cumulative regret minimization and (fixed-confidence) BAI with minimal samples. Although these problems are typically analyzed within the same bandit framework using exponential family distributions, their fundamentally different goals lead to distinct methodological approaches. A detailed comparative discussion can be found in Section 6.*

Exponential families. The canonical single-parameter exponential families of distributions, which encompass a wide range of common distributions such as Gaussian (with fixed variance), Bernoulli, exponential, and Gamma (with fixed shape parameter), have been widely embraced in the bandit literature (Agrawal, 1995; Cappé et al., 2013; Korda et al., 2013; Garivier and Kaufmann, 2016). The distribution ν_θ of one such family is absolutely continuous with respect to some reference measure (e.g., the Lebesgue measure) ρ on $\mathbb{R}$, with the density function

$$\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) = \exp(\theta x - b(\theta)),$$

where the log-partition function $b(\theta) = \log\left(\int_{\mathbb{R}} \exp(\theta x)\,\mathrm{d}\rho(x)\right)$ is infinitely differentiable, and the parameter $\theta \in \Theta := \{\theta \in \mathbb{R} : b(\theta) < \infty\}$. Then it can be verified that the mean and variance of ν_θ are equal to $b'(\theta)$ and $b''(\theta) > 0$, respectively. Therefore, the mapping between the parameter θ and the mean value of ν_θ is one-to-one. In other words, a distribution ν_θ of the family can be uniquely characterized by its mean. Let $I \subseteq \mathbb{R}$ represent the collection of means of ν_θ for all $\theta \in \Theta$. Furthermore, we assume that the variances in the exponential family are bounded by a constant $V > 0$.[2] In this work, the concept of Kullback–Leibler (KL) divergence is employed extensively. For two distributions ν_θ and $\nu_{\theta'}$ with means μ and μ', their KL divergence is denoted as $\mathrm{kl}(\mu, \mu')$ and is explicitly given by:

$$\mathrm{kl}\left(\mu, \mu'\right) = b(\theta') - b(\theta) - b'(\theta)(\theta' - \theta).$$

2. This assumption can be relaxed by assuming that V is an upper bound on the variance of reward distributions with means in the interval $[\min_{i\in[K]} \mu_i, \max_{i\in[K]} \mu_i]$. This is achieved by replacing the maximal inequality in Lemma 9 with the adaptive version provided in Lemma B.5 of Jin et al. (2024). For clarity of exposition, we assume bounded variances within the exponential family.

Several straightforward yet crucial properties include that $\mathrm{kl}(\mu, \mu') = 0$ if and only if $\mu = \mu'$, and with μ fixed, $\mathrm{kl}(\mu, \mu')$ is monotonically increasing in μ' for all $\mu' \geq \mu$. Additional useful properties of the KL divergence between exponential family distributions are detailed in Appendix A. For a comprehensive introduction to exponential families, we refer to Lehmann and Casella (2006).

Other notations. Let $\mathcal{P}_K := \{x \in [0,1]^K : \|x\|_1 = 1\}$ denote the probability simplex in $\mathbb{R}^K$. For any $\mu, \mu' \in (0,1)$, the KL divergence between two Bernoulli distributions with means μ and μ' is denoted as $\mathrm{kl}_{\mathcal{B}}(\mu, \mu') := \mu \log(\mu/\mu') + (1-\mu)\log((1-\mu)/(1-\mu'))$. For each arm $i \in [K]$, let $N_i(t) := \sum_{s=1}^{t} \mathbb{1}\{A_s = i\}$ and $\hat{\mu}_i(t) := \sum_{s=1}^{t} X_s \mathbb{1}\{A_s = i\}/N_i(t)$ denote its total number of pulls and empirical estimate of the mean up to time t, respectively. Additionally, throughout this paper, we adopt standard asymptotic notations, including little o, big O, and little omega, always with respect to the confidence level δ (tending to zero). Specifically, $f(\delta) = \omega(g(\delta))$ means that f grows strictly faster than g as $\delta \to 0^+$, i.e., $\liminf_{\delta \to 0^+} |f(\delta)/g(\delta)| = \infty$.

3. Lower Bound

In this section, we explore the fundamental limits for the problem of best arm identification (BAI) with minimal regret. Specifically, we establish an instance-dependent information-theoretic lower bound on the expected cumulative regret $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]$ for all δ-PAC BAI algorithms. Furthermore, we derive an impossibility result regarding the expected sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ of any *asymptotically optimal* BAI algorithm. These findings provide crucial insights for the design of our algorithm, which will be introduced in Section 4.

We first state the lower bound on the expected cumulative regret $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]$ in the following.

Theorem 3 (Information-theoretic lower bound) *For a fixed confidence level $\delta \in (0,1)$ and instance $\boldsymbol{\mu} \in \mathcal{M}$, any δ-PAC BAI algorithm satisfies that*

$$\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] \geq \mathrm{I}^*(\boldsymbol{\mu}) \mathrm{kl}_{\mathcal{B}}(\delta, 1-\delta)$$

where

$$\mathrm{I}^*(\boldsymbol{\mu}) := \sum_{i=2}^{K} \frac{\Delta_i}{\mathrm{kl}\left(\mu_i, \mu_1\right)}. \tag{2}$$

Furthermore,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]}{\log(1/\delta)} \geq \mathrm{I}^*(\boldsymbol{\mu}). \tag{3}$$

The proof of Theorem 3 is deferred to Appendix B, which employs the ubiquitous change-of-measure argument, dating back to Chernoff (1959), along with some simplifications using optimization techniques.

We refer to $\mathrm{I}^*(\boldsymbol{\mu})$ as the *hardness parameter* for the problem of BAI with minimal regret. This quantity is a function of the suboptimality gaps and the KL divergences between arm i and the best arm for all suboptimal arms $i > 1$. In the asymptotic scenario where the confidence level δ tends to zero, the expected cumulative regret $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]$ is lower bounded

by $\mathrm{I}^*(\boldsymbol{\mu})\log(1/\delta)$. It is worth noting that this result is tight in view of the theoretical performance of our algorithm DKL-UCB presented in Section 4. To formalize this, we introduce the concept of asymptotic optimality as follows.

Definition 4 (Asymptotic optimality) *For the problem of BAI with minimal regret, a δ-PAC algorithm is said to be asymptotically optimal if, for all bandit instances $\boldsymbol{\mu} \in \mathcal{M}$, its expected cumulative regret matches the lower bound asymptotically, i.e.,*

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

Apparently, the cumulative regret of any asymptotically optimal BAI algorithm is on the order of $\Theta(\log(1/\delta))$. One may wonder whether there exist other common properties shared by all asymptotically optimal algorithms. In particular, is it possible for the sample complexity (stopping time) of an asymptotically optimal algorithm to attain the same order of $\Theta(\log(1/\delta))$? We answer this question in the negative through the subsequent impossibility result.

Theorem 5 (Impossibility result) *For the problem of BAI with minimal regret, any asymptotically optimal δ-PAC algorithm satisfies*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \omega(\log(1/\delta))$$

for all bandit instances $\boldsymbol{\mu} \in \mathcal{M}$.

Theorem 5 demonstrates that the pursuit of identifying the optimal arm while minimizing the expected cumulative regret necessarily incurs a sample complexity on the order of $\omega(\log(1/\delta))$. This phenomenon can be understood through the inherent tension between learning the best arm and regret minimization in sequential decision-making.

On one hand, to reliably identify the best arm, the agent must gather sufficient statistical evidence to distinguish it from suboptimal alternatives. This fundamentally requires sampling all arms, including suboptimal ones, a number of times proportional to $\log(1/\delta)$, in order to accumulate enough information to rule out incorrect arms. On the other hand, minimizing cumulative regret incentivizes the agent to pull suboptimal arms as infrequently as possible, and instead exploit the arm that appears to be the best.

Since pulling the optimal arm incurs zero regret, it is appealing for the agent to select it as frequently as possible to refine its estimate of the optimal mean. However, this creates an inefficiency from the perspective of BAI: repeatedly sampling the best arm yields diminishing returns in distinguishing it from competing arms, as the critical information lies in *comparisons* with suboptimal arms. As a result, even though the number of pulls of suboptimal arms is kept at the minimal logarithmic order, the agent must compensate by sampling the best arm significantly more often to satisfy the confidence requirement. This leads to a total sample complexity that is strictly larger than $\Theta(\log(1/\delta))$.

Importantly, the agent does not know the identity of the optimal arm a priori and must learn it adaptively from data. Thus, it is forced to operate under uncertainty, balancing exploration and exploitation without oracle knowledge. Theorem 5 formalizes the consequence of this tradeoff: any algorithm that achieves asymptotically optimal regret must necessarily incur a super-logarithmic number of total samples.

We now provide a slightly more technical proof sketch of Theorem 5, deferring the details to Appendix C.

Proof sketch of Theorem 5. Assume there exists an asymptotically optimal δ-PAC algorithm with $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = O(\log(1/\delta))$. Then, by asymptotic optimality (Definition 4),

$$\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] = \sum_{i>1} \Delta_i \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \sim \mathrm{I}^*(\boldsymbol{\mu}) \log(1/\delta),$$

which implies

$$\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \sim \frac{\log(1/\delta)}{\mathrm{kl}(\mu_i, \mu_1)}, \quad \forall\, i > 1.$$

Thus, suboptimal arms are sampled $\Theta(\log(1/\delta))$ times.

By the change-of-measure argument (Kaufmann et al., 2016, Lemma 1),

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \cdot \mathrm{kl}(\mu_i, \lambda_i) \geq \mathrm{kl}_{\mathcal{B}}(\delta, 1-\delta), \tag{4}$$

where $\mathrm{Alt}(\boldsymbol{\mu})$ is the set of alternative instances where arm 1 is not the best arm. Fixing $j > 1$ and restricting the infimum to alternative instances where only arms 1 and j are modified, we define the function $\Phi(x, y) := \inf_{\lambda \in I} \big(x\,\mathrm{kl}(\mu_1, \lambda) + y\,\mathrm{kl}(\mu_j, \lambda)\big)$. Through the simplification of the infimum in (4), we obtain

$$\Phi\Bigg(\underbrace{\frac{\mathbb{E}_\mu[N_1(\tau_\delta)]}{\log(1/\delta)}}_{=:\,x}, \underbrace{\frac{1}{\mathrm{kl}(\mu_j \mu_1)}}_{=:\,y} \Bigg) \geq 1.$$

However, for any finite x, we have $\Phi(x, y) < y\mathrm{kl}(\mu_j, \mu_1) = 1$, which forces $x \to +\infty$. Hence $\mathbb{E}_\mu[N_1(\tau_\delta)] = \omega(\log(1/\delta))$, implying $\mathbb{E}_\mu[\tau_\delta] = \omega(\log(1/\delta))$.

4. The Double KL-UCB Algorithm

In this section, we introduce a conceptually simple algorithm, namely Double KL-UCB (or DKL-UCB), to identify the best arm with minimal regret. The pseudocode for DKL-UCB is presented in Algorithm 1 and further elucidated in the subsequent discussion.

As its name suggests, our algorithm DKL-UCB leverages two types of Kullback–Leibler upper confidence bounds (UCBs) to guide arm sampling decisions. During the initialization phase, each arm is sampled exactly once. Subsequently, at each time t, we compute two UCBs, $U_i^f(t)$ and $U_i^g(t)$, for each arm $i \in [K]$. These UCBs, defined in Equations (5) and (6), are indexed by exploration functions:

$$f(t) = 3 \log t \quad \text{and} \quad g(\delta, t) = \log\left(\frac{2Kt^2}{\delta}\right).$$

For ease of presentation, we refer to $U_i^f(t)$ and $U_i^g(t)$ as f-UCB and g-UCB, respectively. In our algorithm as well as its analysis, two candidate arms are of particular significance: let A_t^f denote the arm with the highest f-UCB, and A_t^g represent the arm with the highest

Algorithm 1 Double KL-UCB (or DKL-UCB)

Input: Arm set $[K]$ and confidence level $\delta \in (0,1)$.

1: Sample each arm once, and set $t = K$.
2: **repeat**
3: Update $\hat{\mu}_i(t)$ and $N_i(t)$ for all $i \in [K]$, and increment $t \leftarrow t+1$.
4: For each arm $i \in [K]$, compute the quantities

$$U_i^f(t) = \sup\left\{\mu \in I : \mathrm{kl}\left(\hat{\mu}_i(t-1), \mu\right) \le \frac{f(t)}{N_i(t-1)}\right\} \tag{5}$$

and

$$U_i^g(t) = \sup\left\{\mu \in I : \mathrm{kl}\left(\hat{\mu}_i(t-1), \mu\right) \le \frac{g(\delta,t)}{N_i(t-1)}\right\}. \tag{6}$$

5: Let

$$A_t^f = \operatorname*{arg\,max}_{i\in[K]} U_i^f(t) \quad \text{and} \quad A_t^g = \operatorname*{arg\,max}_{i\in[K]\setminus\{A_t^f\}} U_i^g(t). \tag{7}$$

6: Flip a coin with bias (probability of heads) $\beta(\delta)$, as defined in (8).
7: If the outcome is heads, then sample $A_t = A_t^f$; otherwise, sample $A_t = A_t^g$.
8: **until** $L^g_{A_t^f}(t) > U^g_{A_t^g}(t)$ ▷ See the definition of $L_i^g(t)$ in (9)

Output: The candidate best arm $i_{\mathrm{out}} = A_t^f$.

g-UCB excluding A_t^f; see Equation (7). We then proceed to sample either A_t^f or A_t^g in a randomized fashion. Specifically, we toss a biased coin, with the probability of landing heads given by

$$\beta(\delta) = 1 - \min\left\{\frac{1}{\log\log(1/\delta)}, \frac{1}{2}\right\}. \tag{8}$$

If the outcome is heads, we pull the arm $A_t = A_t^f$; otherwise, we pull $A_t = A_t^g$.

Due to its objective of best arm identification in the fixed-confidence setting, our algorithm needs to actively stop when the confidence level δ is reached. For the stopping rule, we utilize a variant of the lower confidence bound, based on the exploration function $g(\delta,t)$. In particular, for each arm $i \in [K]$, we define its g-LCB as follows:

$$L_i^g(t) = \inf\left\{\mu \in I : \mathrm{kl}\left(\hat{\mu}_i(t-1), \mu\right) \le \frac{g(\delta,t)}{N_i(t-1)}\right\}. \tag{9}$$

Ultimately, our algorithm terminates when the g-LCB of A_t^f exceeds the g-UCB of A_t^g. In other words, it stops at the stopping time

$$\tau_\delta = \inf\left\{t \in \mathbb{N} : L^g_{A_t^f}(t) > U^g_{A_t^g}(t)\right\}, \tag{10}$$

and recommends $A^f_{\tau_\delta}$ as the identified best arm i_{out}.

Remark 6 *It is worth discussing some similarities between our algorithm* DKL-UCB *and other algorithms in the multi-armed bandit literature. If our algorithm consistently opts to pull the arm A_t^f, then its sampling rule mirrors that of* KL-UCB*, a celebrated algorithm for cumulative regret minimization (Cappé et al., 2013), up to the choice of the exploration function $f(t)$.*

*Moreover, concerning the randomized dynamics within the sampling rule, our algorithm shares commonalities with Top-Two Thompson Sampling (*TTTS*), which is designed for fixed-confidence best arm identification (Russo, 2020). Specifically, both algorithms incorporate randomness in selecting one of two candidate arms. However, apart from how the two candidates (called leader and challenger in the* TTTS *literature) are determined, another significant distinction lies in the selection probability. In* TTTS*, each candidate is pulled with a fixed probability, which is independent of the confidence level δ. In contrast, in our algorithm, as δ approaches zero, the coin bias $\beta(\delta)$ tends toward one. Consequently, the arm A_t^f dominates the proportion of pulls. In this respect, the alternative candidate A_t^g serves as a minor yet critical supplement to the arm sampling rule.*

5. Theoretical Analysis of DKL-UCB

In this section, we theoretically analyze the performance of our algorithm DKL-UCB from multiple perspectives. Our main results are presented in Theorem 7 below. The complete proof of Theorem 7 are deferred to Appendix D. In Section 5.2, we discuss the primary technical challenges and provide a proof sketch of Theorem 7.

5.1 Main Results

Theorem 7 *For every confidence level $\delta \in (0,1)$, the* DKL-UCB *algorithm (Algorithm 1) has the following properties. For all bandit instances $\boldsymbol{\mu} \in \mathcal{M}$, it guarantees that*

$$\mathbb{P}(i_{\text{out}} \neq 1) \leq \delta. \qquad \text{(DKL-UCB is } \delta\text{-PAC)}$$

Furthermore, when $\delta \to 0$, the regret of DKL-UCB *satisfies*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]}{\log(1/\delta)} \leq \mathrm{I}^*(\boldsymbol{\mu}), \qquad \text{(Regret Bound)}$$

and

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta) \cdot (\log\log(1/\delta))^2} = 0. \qquad \text{(Sample Complexity)}$$

Theorem 7 first guarantees the correctness of our algorithm: the probability of recommending a suboptimal arm is no more than δ. As for the expected cumulative regret, by comparing the instance-dependent upper bound in Theorem 7 with the corresponding lower bound in Theorem 3, it is evident that our algorithm DKL-UCB is asymptotically optimal for the problem of BAI with minimal regret. Particularly, for all bandit instances $\boldsymbol{\mu} \in \mathcal{M}$, the expected regret of our algorithm satisfies the following limiting behaviour:

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]}{\log(1/\delta)} = \mathrm{I}^*(\boldsymbol{\mu}).$$

Furthermore, regarding the expected sample complexity, in conjunction with the impossibility result in Theorem 5, Theorem 7 shows that our algorithm satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \omega(\log(1/\delta)) \cap o\left(\log(1/\delta) \cdot (\log\log(1/\delta))^2\right). \tag{11}$$

Even though the principal performance metric for our problem is not the expected sample complexity, the result in (11) demonstrates that the sample complexity of our algorithm DKL-UCB is close to the fundamental barrier in Theorem 5, up to a small multiplicative factor of the order of $(\log\log(1/\delta))^2$. Consequently, our algorithm is nearly optimal in terms of sample complexity while achieving optimality in the cumulative regret.

5.2 Technical Challenges and Proof Outline

In this section, we discuss the key technical challenges and outline the proof of Theorem 7. Readers primarily interested in the theoretical and experimental results may choose to skip this subsection on their first reading and proceed directly to Section 6.

Our instance-dependent lower bound in Theorem 3 suggests that the optimal regret upper bound could be $\mathrm{I}^*(\boldsymbol{\mu})\log(1/\delta) + o(\log(1/\delta))$ as the confidence level δ tends to zero. In contrast, the optimal regret bound for cumulative regret minimization is $\mathrm{I}^*(\boldsymbol{\mu})\log(T) + o(\log T)$ as the time horizon T tends to infinity, which indicates that an optimal regret minimization algorithm pulls each suboptimal arm $O(\log T)$ times over time T (see (12) in Section 6). Thus, one natural strategy for our problem is to run the KL-UCB algorithm and check whether the best arm can be reliably identified over time. However, the sample complexity of such an algorithm is easily seen to be $\Theta(1/\delta)$, which is significantly greater than $\Theta(\log(1/\delta))$.

To address this problem, we introduce DKL-UCB, which utilizes *two* UCB indices, f-UCB and g-UCB, as defined in (5) and (6), respectively. In our approach, f-UCB, with $f(t) = 3\log t$, is primarily employed for exploring the best arm. This choice differs from the standard KL-UCB algorithm (Cappé et al., 2013) where the exploration function $f(t)$ is approximately $\log t$. While one might suspect that $f(t) = 3\log t$ could result in suboptimal asymptotic regret for the regret minimization task due to the potential non-optimality of the constant 3, the inflated choice of $f(t)$ does not lead to suboptimal regret for our purpose. Roughly speaking, this is because the stopping time τ_δ satisfies $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = O(\log^2(1/\delta))$, and hence the regret of f-UCB (when $A_t = A_t^f$) within $O(\log^2(1/\delta))$ time steps is of order $o(\log(1/\delta))$. Therefore, the inflated choice of $f(t)$ does not hinder the asymptotic optimality properties of DKL-UCB for the unique problem under consideration.

In our analytical framework, we partition the time horizon into intervals of exponentially growing lengths; see Figure 1 for an illustration. Let $\gamma = \log\log(1/\delta)$, $\epsilon = \gamma^{-1/4}$ and $h(\delta) = \log(1/\delta) \cdot \gamma/\epsilon = \log(1/\delta) \cdot (\log\log(1/\delta))^{5/4}$. Then the partition is defined by the time points $t_r = 2^r h(\delta)$, with each subinterval taking the form $(t_r, t_{r+1}]$, and an additional initial interval $(0, t_0]$. Since the best arm is pulled for a sufficient number of times through f-UCB, we can show that $\mathbb{P}(L_1^g(t_r) \geq \mu_1 - \epsilon) \geq 1 - 1/t_r^2$. In view of this, to check that the stopping rule in (10) is fulfilled, it suffices to additionally check that, with high probability, $U_i^g(t_r) < \mu_1 - \epsilon$ for all suboptimal arms $i \in [K] \setminus \{1\}$.

To speed up the exploration of suboptimal arms, we incorporate supplementary randomized exploration using g-UCB, which is specifically tailored for the suboptimal arms.

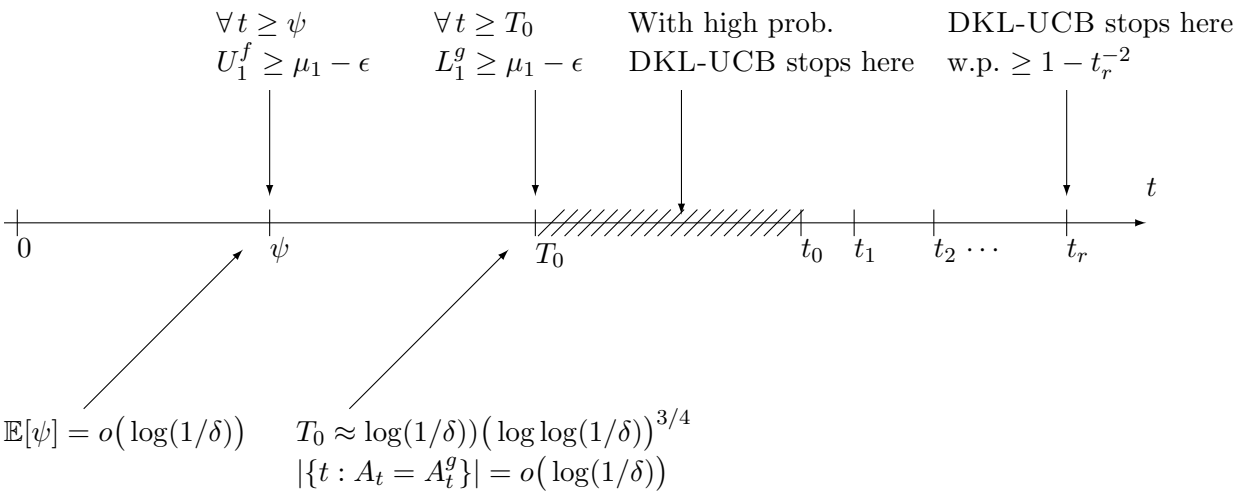


Figure 1: Illustration of the proof.

Notably, the exploration function $g(\delta, t)$ associated with g-UCB depends on the confidence level δ. In our algorithm, we exclude A_t^f, the arm with the highest f-UCB, when determining A_t^g, and pull the arm A_t^g with a probability of $1 - \beta(\delta) = \min\{1/\gamma, 1/2\}$. At time t_r, the expected number of times that we pull the arm A_t^g is $t_r(1 - \beta(\delta)) = \omega(2^r \log(1/\delta))$, which ensures that $\max_{i>1} U_i^g(t_r) < \mu_1 - \epsilon$ with probability at least $1 - 1/t_r^2$. Furthermore, given that $L_1^g(t_r) \geq \mu_1 - \epsilon$, the probability that the algorithm terminates after t_r is at most $1/t_r^2$. This proposition is formally established in Lemma 15. In its proof detailed in Appendix E.1, we introduce four events that are *a priori* dependent, but we meticulously decouple their complex interdependencies. Consequently, the regret incurred from pulling arms beyond time t_0 can be bounded such that optimality is ensured.

Here, we provide further insights into the role of the parameter $\beta(\delta)$. This parameter is crucial in accelerating the learning process and thereby reducing the sample complexity. As δ approaches 0, $\beta(\delta)$ approaches 1, indicating that the algorithm DKL-UCB increasingly favors A_t^f over A_t^g. This aligns with the lower bound intuition, where: (1) to achieve asymptotic optimality, the optimal arm must be selected $\omega(\log(1/\delta))$ times; and (2) for large values of t, when $A_t = A_t^f$, the algorithm primarily chooses the optimal arm. Conversely, with probability $1-\beta(\delta)$, the algorithm selects $A_t = A_t^g$, which is likely a suboptimal choice for large t. Therefore, if t exceeds $\log(1/\delta) \cdot (\log\log(1/\delta))^2$, DKL-UCB will, with high probability, select suboptimal arms at least $\log(1/\delta) \cdot (\log\log(1/\delta))^2 \cdot (1-\beta(\delta)) = \omega(\log(1/\delta))$ times. This guarantees sufficient exploration of suboptimal arms before reaching the time $\log(1/\delta) \cdot (\log\log(1/\delta))^2$.

To bound the total expected cumulative regret, a key theoretical challenge arises from the regret incurred before t_0, i.e., $\mathbb{E}_{\boldsymbol{\mu}}[R(t_0)]$. Since $t_0 = \omega(\log(1/\delta))$, the regret bound of $\mathbb{E}_{\boldsymbol{\mu}}[R(t_0)]$ could be suboptimal if not appropriately managed. Let ψ represent the smallest time step t such that $U_1^f(t) \geq \mu_1 - \epsilon$, i.e., $\psi := \inf\{t : U_1^f(t) \geq \mu_1 - \epsilon\}$. By leveraging a novel concentration inequality for exponential families (Lemma 11 in Appendix D), we can establish that $\mathbb{E}_{\boldsymbol{\mu}}[\psi] = O(1/\epsilon^2) = o(\log(1/\delta))$.

Our analysis relies on a carefully chosen time step T_0, ensuring that, with high probability, the optimal arm is well-estimated at time T_0 (i.e., $L_1^g(T_0) \geq \mu_1 - \epsilon$) and the regret from suboptimal arms remains within an acceptable range. Specifically, we define T_0 as follows:

$$T_0 = \min \left\{ \psi + \underbrace{\sum_{i=2}^{K} \sum_{t=\psi+1}^{t_0} \mathbf{1} \left\{ A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) \leq \frac{f(t_0)}{N_i(t-1)} \right\}}_{Q} + \gamma\epsilon \log \frac{1}{\delta},\, t_0 \right\}.$$

Here we consider the nontrivial case where $T_0 < t_0$. Since T_0 exceeds ψ, we have $U_1^f(T_0) > \mu_1 - \epsilon$. Besides, according to the dynamics of our algorithm, the term Q represents an upper bound on the number of pulls of the suboptimal arms through f-UCB. That is, Q bounds the number of times $A_t = A_t^f \in [K] \setminus \{1\}$ for t between $\psi + 1$ and t_0. Therefore, the $\gamma\epsilon \log(1/\delta)$ time steps within the definition of T_0 encompass either pulls of the optimal arm or the time steps when $A_t = A_t^g$. Utilizing a concentration bound, we can deduce that the number of times when $A_t = A_t^g$ within the $\gamma\epsilon \log(1/\delta)$ time steps is at most $2\epsilon \log(1/\delta)$, with high probability. As a result, by time T_0, the algorithm will have pulled the optimal arm a sufficient number of times such that $L_1^g(T_0) \geq \mu_1 - \epsilon$ with probability at least $1 - 1/t_0$.

Finally, the number of pulls of the suboptimal arms before time T_0 is of order $O(\psi + Q + \epsilon \log(1/\delta)) = o(\log(1/\delta))$, which does not result in suboptimal regret. Additionally, Lemma 14 demonstrates that from $T_0 + 1$ to t_0, the number of pulls of any suboptimal arm $i \in [K] \setminus \{1\}$ does not exceed $g(\delta, t_0)/\mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon)$ with a probability of at least $1 - 1/t_0$. Altogether, we can bound the regret incurred before t_0 optimally.

6. Discussion: Comparisons to Related Problems

In this following, we compare our investigated problem—best arm identification with minimal regret—with two other extensively studied problems in the bandit literature, focusing on their respective asymptotic performances.

Cumulative regret minimization. The objective of this problem is to maximize the expected cumulative rewards up to the time horizon $T \in \mathbb{N}$, which might be either known or unknown to the agent a priori. Equivalently, the agent aims at minimizing the expected cumulative regret $\mathbb{E}_{\boldsymbol{\mu}}[R(T)]$.

In the seminal work by Lai and Robbins (1985), it was proven that any *consistent*[3] regret minimization algorithm must satisfy that for all bandit instances $\boldsymbol{\mu} \in \mathcal{M}$,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R(T)]}{\log T} \geq \mathrm{I}^*(\boldsymbol{\mu}), \tag{12}$$

where $\mathrm{I}^*(\boldsymbol{\mu})$ is the same as the quantity defined in Equation (2).

Although BAI with minimal regret and cumulative regret minimization represent distinctly different tasks, the coincidence regarding $\mathrm{I}^*(\boldsymbol{\mu})$ suggests profound connections between them. Specifically, the asymptotic lower bound presented in (3) for BAI with minimal

3. A regret minimization algorithm is said to be consistent, if for all bandit instances $\boldsymbol{\mu} \in \mathcal{M}$ and all $\alpha > 0$, $\mathbb{E}_{\boldsymbol{\mu}}[R(T)] = o(T^\alpha)$.

regret follows from the δ-PAC nature of the algorithm, while the asymptotic lower bound in (12) relies on the assumption of consistency. This coincidence might be attributed to the shared consideration of cumulative regret in both problems, and represents an optimal equilibrium of regret among suboptimal arms. We defer a more in-depth investigation into this intriguing phenomenon as future work. In the current study, this insightful finding is leveraged in the design of our algorithm; see Section 5.2.

Best arm identification with minimal samples. In contrast to our problem setup, this problem shares the common goal of identifying the best arm with a prescribed confidence level $\delta \in (0,1)$, but focuses on a distinct performance metric—the expected sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$. Using our notation, the problem can be formally expressed as follows:

$$\begin{aligned} \min \quad & \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \\ \text{s.t.} \quad & \mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty) = 1 \text{ and } \mathbb{P}_{\boldsymbol{\mu}}(i_{\text{out}} \neq 1) \leq \delta. \end{aligned} \tag{13}$$

It is worth highlighting that the fundamental distinction between (1) and (13) lies solely in the choice of the objective function.

The instance-dependent lower bound for the problem of BAI with minimal samples can be characterized through a max-min optimization problem (Garivier and Kaufmann, 2016). Specifically, let $\text{Alt}(\boldsymbol{\mu})$ denote the set of alternative bandit instances where arm 1 is not the best arm. Then for any δ-PAC BAI algorithm, it holds that

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq \Gamma^*(\boldsymbol{\mu})$$

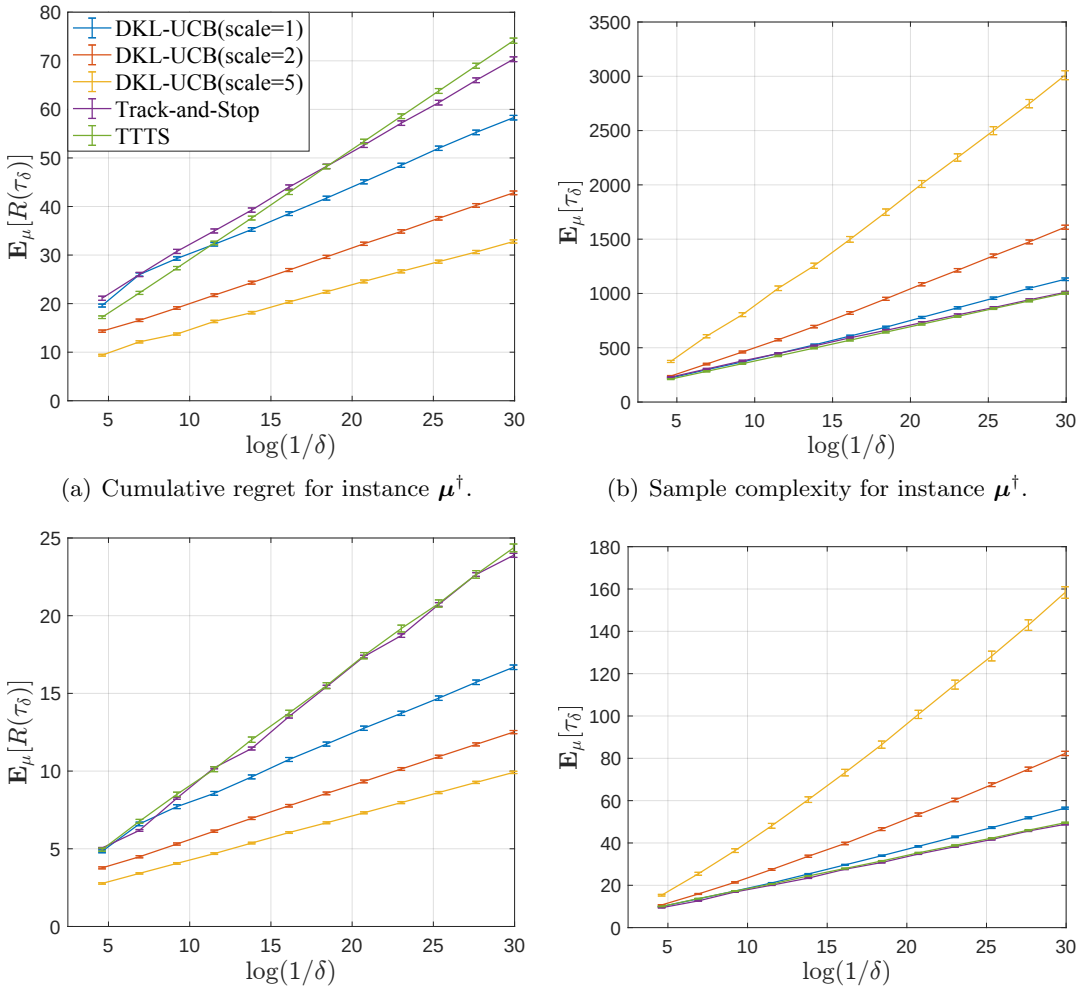
where

$$\Gamma^*(\boldsymbol{\mu})^{-1} := \sup_{w \in \mathcal{P}_K} \inf_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \left(\sum_{i=1}^K w_i \text{kl}\left(\mu_i, \lambda_i\right) \right). \tag{14}$$

Note that this lower bound can be attained asymptotically by approximating the optimal proportion of arm pulls, as indicated by the outer supremum over $w \in \mathcal{P}_K$ in Equation (14) (Garivier and Kaufmann, 2016; Degenne et al., 2019a; Mukherjee and Tajer, 2023).

A natural question emerges when we consider these two problems in further depth: Does an optimal policy for BAI with minimal samples inherently lead to a commensurately low cumulative regret? The following example provides a counterexample to this assertion, underscoring the significance of carefully investigating the problem of BAI with minimal regret.

Example 1 *Consider a two-armed Bernoulli bandit instance $\boldsymbol{\mu} = (1-\mu, \mu)$ with $\mu \in (0, 1/2)$. Here, as with the rest of the paper, we focus on the asymptotic regime that δ tends to zero. For BAI with minimal samples, a simple mathematical derivation reveals that the optimal proportion of arm pulls is uniform (see Appendix F), and hence the expected cumulative regret under this scenario is approximately $\frac{(1-2\mu)\log(1/\delta)}{2\text{kl}_{\mathcal{B}}(\mu,1/2)}$. On the other hand, for the optimal policy of BAI with minimal regret, its expected cumulative regret is given by $\frac{(1-2\mu)\log(1/\delta)}{\text{kl}_{\mathcal{B}}(\mu,1-\mu)}$. As the parameter μ approaches zero, the ratio between these two regret values, $\frac{\text{kl}_{\mathcal{B}}(\mu,1-\mu)}{2\text{kl}_{\mathcal{B}}(\mu,1/2)} \simeq \frac{\log(1/\mu)}{2\log 2}$, can become arbitrarily large.*



(a) Cumulative regret for instance $\boldsymbol{\mu}^\dagger$.

(b) Sample complexity for instance $\boldsymbol{\mu}^\dagger$.

(c) Cumulative regret for instance $\boldsymbol{\mu}^\ddagger$.

(d) Sample complexity for instance $\boldsymbol{\mu}^\ddagger$.

Figure 2: Empirical cumulative regrets (up to stopping times) and sample complexities for different confidence levels δ.

7. Numerical Experiments

To validate our theoretical results, we compare our DKL-UCB algorithm with two established approaches: Track-and-Stop (Garivier and Kaufmann, 2016) and Top-Two Thompson Sampling (TTTS) (Russo, 2020). To guarantee fair comparisons, all algorithms employ the generalized likelihood ratio test (GLRT) stopping rule introduced in Garivier and Kaufmann (2016). We note that standard KL-UCB (Garivier and Cappé, 2011; Cappé et al., 2013) never terminates under this stopping rule and is therefore excluded from the comparison.[4]

4. Specifically, the threshold function used in the GLRT stopping rule grows too rapidly for standard KL-UCB to satisfy the stopping criterion.

In our experiments, we consider a family of DKL-UCB algorithms parameterized by a scaling factor $c > 0$ within the definition of $\beta(\delta)$ in (8). Specifically, we modify $\beta(\delta)$ to:

$$\beta_c(\delta) = 1 - \min\left\{\frac{1}{c\ \log\log(1/\delta)}, \frac{1}{2}\right\}. \tag{15}$$

When $c = 1$, $\beta_c(\delta)$ simplifies to the original form of $\beta(\delta)$ given in (8). As $c \to +\infty$, $\beta_c(\delta) \to 1^-$ and the arm A_t^f, defined in (7), is pulled more frequently, making the algorithm increasingly resemble KL-UCB as mentioned in Remark 6. We note, however, that regardless of the exact value of $c > 0$, our proposed DKL-UCB algorithm possesses the same theoretical asymptotic guarantees as described in Theorem 7.

We consider two problem instances: a five-armed and a two-armed Bernoulli bandit with mean vectors

$$\boldsymbol{\mu}^\dagger = [1, 0.9, 0.8, 0.7, 0.6] \quad \text{and} \quad \boldsymbol{\mu}^\ddagger = [0.99, 0.01].$$

The latter instance (i.e., $\boldsymbol{\mu}^\ddagger$) corresponds to the setting described in Example 1 where $\mu = 0.01$ is small. This configuration is expected to highlight a divergence in the performances of DKL-UCB and Track-and-Stop. For each setting, the results are averaged over 1000 independent trials with standard errors displayed as error bars. Our results are depicted in Figure 2. The top and bottom rows display results for instances $\boldsymbol{\mu}^\dagger$ and $\boldsymbol{\mu}^\ddagger$, respectively, while the left and right columns present empirical regrets and sample complexities.

From Figures 2(a) and (c), we observe that the regrets of DKL-UCB are often smaller compared to those of Track-and-Stop and TTTS. This outcome aligns with expectations, as the latter algorithms are specifically designed to achieve asymptotically optimal sample complexity for fixed-confidence BAI and not to minimize cumulative regret. When considering larger values of δ, DKL-UCB with $c = 1$ does not consistently outperform its competitors. However, this limitation can be addressed by selecting a larger scaling factor (such as $c = 5$), indicating that for moderate values of δ, a more aggressive sampling approach of A_t^f relative to A_t^g is beneficial for reducing regret. In our main theorem, we show that as $\delta \to 0$, choosing $\beta(\delta) \to 1$ ensures asymptotic optimality of our algorithm. For finite δ, our experimental results suggest that this choice may also lead to good finite-time regret performance. We leave a more thorough investigation of this for future work. Furthermore, as seen in Figure 2(c), the slope of DKL-UCB (across various scaling factors) is considerably smaller than that of Track-and-Stop and TTTS. This observation supports our theoretical finding in Example 1, highlighting the significant regret difference between DKL-UCB and algorithms designed for BAI with minimal samples.

Conversely, Figures 2(b) and (d) show that Track-and-Stop and TTTS outperform DKL-UCB in terms of sample complexity. This is also expected, as these algorithms are specifically tailored for efficient BAI with minimal samples.

Overall, the experimental results corroborates our theoretical findings: DKL-UCB excels at BAI with minimal regret. It outperforms baseline algorithms not optimized for this balancing act, highlighting the inherent trade-off between minimizing regret and sample complexity in the study of multi-armed bandits.

8. Conclusions and Future Work

In this paper, we investigate the problem of fixed-confidence best arm identification while minimizing the expected cumulative regret. By analyzing the instance-dependent lower bound, we clearly demonstrate the differences between this problem and other problems in the bandit literature. Concurrently, we design and analyze Double KL-UCB (or DKL-UCB), which stands out for its utilization of dual upper confidence bounds. Our rigorous theoretical examination firmly confirms its asymptotic optimality in terms of cumulative regret as well as its near-optimality in sample complexity.

While our algorithm DKL-UCB is applicable for any fixed confidence level $\delta \in (0,1)$, its theoretical guarantees are asymptotic. In our analysis, we leverage asymptotic techniques to streamline the proof process and presentation. Although non-asymptotic results are possible with our current proof strategies, they would be considerably more complex. We leave a more concise non-asymptotic performance analysis for future work.

Moreover, a fruitful direction for future research lies in establishing bounds for the worst case regret for our problem. Note that for the problem of cumulative regret minimization, the two predominant performance metrics are asymptotic and worst case regrets. In particular, the optimal bound for the latter is $O(\sqrt{KT})$. While our work studies the asymptotic regret performance for fixed-confidence BAI, the exploration of worst case regret remains open.

An intriguing open question arises regarding the fundamental limits of sample complexity, denoted as $\mathbb{E}[\tau_\delta]$, as δ approaches zero. From Definition 4, we understand the concept of asymptotic optimality and, as per Theorem 5, any asymptotic optimal algorithm inherently incurs a sample complexity $\mathbb{E}[\tau_\delta] = \omega\big(\log(1/\delta)\big)$ as $\delta \to 0$. This concept can be extended to ζ-asymptotic optimality. An algorithm is said to be ζ-*asymptotically optimal* for $\zeta > 1$ if:

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu[R(\tau_\delta)]}{\log(1/\delta)} \leq \zeta \sum_{i \in [K] \setminus \{1\}} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu_1)}. \tag{16}$$

The question is to explore the fundamental limits of $\mathbb{E}[\tau_\delta]$ as a function of ζ as δ vanishes under the constraint in (16). Specifically, does $\mathbb{E}[\tau_\delta]$ scale as $\Theta\big(\log(1/\delta)\big)$, and if it does, what is the exact constant involved? This remains an open question for future research.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. T. Jin and V. Y. F. Tan are supported by a Singapore Ministry of Education (MOE) AcRF Tier 2 grant under grant number A-8004062-00-00.

Appendix A. Auxiliary Results

A.1 Exponential Family Distributions

For exponential families of distribution, the Kullback–Leibler (KL) divergence can be equivalently expressed as the Bregman divergence between the parameters or the means. Specif-

ically, for two distributions ν_θ and $\nu_{\theta'}$ with means μ and μ', it holds that

$$\begin{aligned}\mathrm{kl}\left(\mu, \mu'\right) &= b\left(\theta'\right) - b(\theta) - b'(\theta)\left(\theta' - \theta\right) \\ &= b^*(\mu) - b^*(\mu') - \left(b'\right)^{-1}(\mu')(\mu - \mu')\end{aligned}$$

where b^* is the convex conjugate of b.

Lemma 8 (Ménard and Garivier (2017)) *For any $\mu, \mu' \in I$, it holds that*

$$\mathrm{kl}\left(\mu, \mu'\right) \geq \frac{(\mu - \mu')^2}{2V}.$$

Lemma 9 (Maximal Inequality (Ménard and Garivier, 2017)) *Let N_1 and N_2 be two real numbers, and $\hat{\mu}_n$ be the empirical mean of n i.i.d. random variables drawn according to some exponential family distribution with mean μ. Let V be the maximum variance of the distribution. Then, for every $x \leq \mu$,*

$$\begin{aligned}&\mathbb{P}(\exists N_1 \leq n \leq N_2, \hat{\mu}_n \leq x) \leq e^{-N_1 \cdot \mathrm{kl}(x,\mu)}, \\ &\mathbb{P}(\exists N_1 \leq n \leq N_2, \hat{\mu}_n \leq x) \leq e^{-N_1 (x-\mu)^2/(2V)}.\end{aligned} \tag{17}$$

Moreover, for every $x \geq \mu$,

$$\mathbb{P}(\exists N_1 \leq n \leq N_2, \hat{\mu}_n \geq x) \leq e^{-N_1 \cdot \mathrm{kl}(x,\mu)}. \tag{18}$$

A.2 Other Technical Lemmas

Lemma 10 *Consider any function $f : [0, \infty)^2 \to \mathbb{R}$. If f is continuous on $[0, \infty)^2$, then for any sequence $(x_n) \subset \mathbb{R}$ and convergent sequence $(y_n) \subset \mathbb{R}$, it holds that*

$$\liminf_{n\to\infty} f\left(x_n, y_n\right) \leq f\left(\liminf_{n\to\infty} x_n, \lim_{n\to\infty} y_n\right).$$

Proof For the sake of brevity, we write $x_0 := \liminf_{n\to\infty} x_n$ and $y_0 := \lim_{n\to\infty} y_n$ (since (y_n) is convergent). By a property of the lim inf, there exists a convergent subsequence (x_{n_k}) of (x_n) such that $\lim_{k\to\infty} x_{n_k} = x_0$. Along this subsequence, we also have $\lim_{k\to\infty} y_{n_k} = y_0$, since every subseqence of a convergent sequence converges to the same limit. By the definition of lim inf,

$$\liminf_{n\to\infty} f(x_n, y_n) = \inf E \tag{19}$$

where E is the set of subsequential limits of the sequence $(f(x_n, y_n))$. Now, note that $(f(x_{n_k}, y_{n_k}))$ is convergent. Indeed, its limit is

$$\lim_{k\to\infty} f(x_{n_k}, y_{n_k}) = f\left(\lim_{k\to\infty} x_{n_k}, \lim_{k\to\infty} y_{n_k}\right) = f(x_0, y_0)$$

where the first equality follows from the continuity of f. As a result, we may upper bound in the infimum in (19) by choosing the convergent subsequence $(f(x_{n_k}, y_{n_k}))$, yielding

$$\liminf_{n\to\infty} f(x_n, y_n) \leq \lim_{k\to\infty} f(x_{n_k}, y_{n_k}) = f(x_0, y_0)$$

as previously established. This completes the proof of Lemma 10. ■

Appendix B. Proof of Theorem 3

Proof For a fixed confidence level $\delta \in (0,1)$ and instance $\boldsymbol{\mu} \in \mathcal{M}$, consider any δ-PAC best arm identification algorithm. Since the hardness parameter $\mathrm{I}^*(\boldsymbol{\mu})$ is finite, the situation that $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]$ is infinite is trivial. Henceforth, we assume that $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]$ is finite.

Recall that $\mathrm{Alt}(\boldsymbol{\mu})$ represents the set of alternative instances such that arm 1 is not the best arm. Consider an arbitrary instance $\boldsymbol{\lambda}$ in $\mathrm{Alt}(\mu)$. By applying the *transportation* inequality (Kaufmann et al., 2016, Lemma 1) and the KL divergence for the underlying exponential family, we have

$$\sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \geq \mathrm{kl}_{\mathcal{B}}(\delta, 1-\delta).$$

Since the above inequality holds for all instances in $\mathrm{Alt}(\boldsymbol{\mu})$, we have

$$\begin{aligned}
\mathrm{kl}_{\mathcal{B}}(\delta, 1-\delta) &\leq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \\
&\leq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \\
&= \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \sum_{i=2}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \\
&= \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] \left(\sum_{i=2}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \right) \\
&= \mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \left(\sum_{i=2}^{K} \frac{\Delta_i \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} \cdot \frac{\mathrm{kl}\left(\mu_i, \lambda_i\right)}{\Delta_i} \right).
\end{aligned} \tag{20}$$

Due to the law of total expectation, the regret $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]$ can be decomposed as follows:

$$\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{\tau_\delta} \Delta_{A_t} \right] = \sum_{i=2}^{K} \Delta_i \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)].$$

As a result, the vector

$$\left(0, \frac{\Delta_2 \mathbb{E}_{\boldsymbol{\mu}}[N_2(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]}, \frac{\Delta_3 \mathbb{E}_{\boldsymbol{\mu}}[N_3(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]}, \ldots, \frac{\Delta_K \mathbb{E}_{\boldsymbol{\mu}}[N_K(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} \right)$$

constitutes a probability distribution in $\mathcal{P}_K$.

Hence, we can derive the following:

$$\begin{aligned}
\mathrm{kl}_{\mathcal{B}}(\delta, 1-\delta) &\leq \mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \left(\sum_{i=2}^{K} \frac{\Delta_i \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} \cdot \frac{\mathrm{kl}\left(\mu_i, \lambda_i\right)}{\Delta_i} \right) \\
&\leq \mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] \sup_{w \in \mathcal{P}_K, w_1 = 0} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \left(\sum_{i=2}^{K} \frac{w_i \mathrm{kl}\left(\mu_i, \lambda_i\right)}{\Delta_i} \right).
\end{aligned} \tag{21}$$

To establish the desired inequality $\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)] \geq \mathrm{I}^*(\boldsymbol{\mu})\mathrm{kl}_{\mathcal{B}}(\delta, 1-\delta)$, it suffices to show

$$\sup_{w\in\mathcal{P}_K, w_1=0} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu}),\lambda_1=\mu_1} \left(\sum_{i=2}^{K} \frac{w_i \mathrm{kl}\left(\mu_i,\lambda_i\right)}{\Delta_i}\right) = \mathrm{I}^*(\boldsymbol{\mu})^{-1}.$$

By decomposing the feasible set (i.e., $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$ with $\lambda_1 = \mu_1$), we can get

$$\begin{aligned}
&\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu}),\lambda_1=\mu_1} \left(\sum_{i=2}^{K} \frac{w_i \mathrm{kl}\left(\mu_i,\lambda_i\right)}{\Delta_i}\right) \\
&= \min_{j>1} \inf_{\boldsymbol{\lambda}:\lambda_1=\mu_1,\lambda_j>\lambda_1} \left(\sum_{i=2}^{K} \frac{w_i \mathrm{kl}\left(\mu_i,\lambda_i\right)}{\Delta_i}\right) \\
&= \min_{j>1} \inf_{\boldsymbol{\lambda}:\lambda_1=\mu_1,\lambda_j>\lambda_1} \frac{w_j \mathrm{kl}\left(\mu_j,\lambda_j\right)}{\Delta_j} \\
&= \min_{j>1} \frac{w_j \mathrm{kl}\left(\mu_j,\mu_1\right)}{\Delta_j}.
\end{aligned}$$

Now consider the outer optimization problem:

$$\sup_{w\in\mathcal{P}_K, w_1=0} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu}),\lambda_1=\mu_1} \left(\sum_{i=2}^{K} \frac{w_i \mathrm{kl}\left(\mu_i,\lambda_i\right)}{\Delta_i}\right) = \sup_{w\in\mathcal{P}_K, w_1=0} \min_{j>1} \frac{w_j \mathrm{kl}\left(\mu_j,\mu_1\right)}{\Delta_j}.$$

Obviously, the supremum (maximum) is attained if and only if the values of $\frac{w_j \mathrm{kl}(\mu_j,\mu_1)}{\Delta_j}$ are the same across all $j > 1$. Since $\{w_j\}_{j=1}^K$ forms a probability distribution, this occurs when

$$w_j = \frac{\Delta_j / \mathrm{kl}\left(\mu_j,\mu_1\right)}{\sum_{i=2}^{K} \Delta_i / \mathrm{kl}\left(\mu_i,\mu_1\right)}$$

for all $j > 1$.

Therefore, we arrive at

$$\sup_{w\in\mathcal{P}_K, w_1=0} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu}),\lambda_1=\mu_1} \left(\sum_{i=2}^{K} \frac{w_i \mathrm{kl}\left(\mu_i,\lambda_i\right)}{\Delta_i}\right) = \left(\sum_{i=2}^{K} \frac{\Delta_i}{\mathrm{kl}\left(\mu_i,\mu_1\right)}\right)^{-1} = \mathrm{I}^*(\boldsymbol{\mu})^{-1}.$$

Finally, since $\lim_{\delta\to 0} \frac{\mathrm{kl}_{\mathcal{B}}(\delta,1-\delta)}{\log(1/\delta)} = 1$, letting $\delta \to 0$ yields

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]}{\log(1/\delta)} \geq \mathrm{I}^*(\boldsymbol{\mu}).$$

This completes the proof of Theorem 3. ■

Appendix C. Proof of Theorem 5

Proof Consider any asymptotically optimal δ-PAC algorithm and bandit instance $\boldsymbol{\mu} \in \mathcal{M}$. By combining the asymptotic lower bound in Theorem 3 with the definition of asymptotic optimality in Definition 4, we obtain

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]}{\log(1/\delta)} = \mathrm{I}^*(\boldsymbol{\mu}). \tag{22}$$

First, consider Inequality (21) in the proof of Theorem 3. By dividing this inequality by $\log(1/\delta)$ and allowing δ to approach zero, we have

$$1 \leq \lim_{\delta \to 0} \mathrm{I}^*(\boldsymbol{\mu}) \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \left(\sum_{i=2}^{K} \frac{\Delta_i \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} \cdot \frac{\mathrm{kl}\left(\mu_i, \lambda_i\right)}{\Delta_i} \right) \leq 1,$$

which leads to

$$\lim_{\delta \to 0} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \lambda_1 = \mu_1} \left(\sum_{i=2}^{K} \frac{\Delta_i \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} \cdot \frac{\mathrm{kl}\left(\mu_i, \lambda_i\right)}{\Delta_i} \right) = \mathrm{I}^*(\boldsymbol{\mu})^{-1}.$$

Based on the analysis in the proof of Theorem 3, the above equation is equivalent to

$$\lim_{\delta \to 0} \min_{i > 1} \frac{\Delta_i \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} \cdot \frac{\mathrm{kl}\left(\mu_i, \mu_1\right)}{\Delta_i} = \mathrm{I}^*(\boldsymbol{\mu})^{-1},$$

which holds if and only if

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[R(\tau_\delta)]} = \frac{1}{\mathrm{I}^*(\boldsymbol{\mu}) \mathrm{kl}\left(\mu_i, \mu_1\right)} \tag{23}$$

for all $i > 1$.

Consequently, by combining (22) and (23), we establish that for all $i > 1$,

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\log(1/\delta)} = \frac{1}{\mathrm{kl}\left(\mu_i, \mu_1\right)}. \tag{24}$$

Next, consider Inequality (20). Similarly, dividing this inequality by $\log(1/\delta)$ and allowing δ to approach zero yields

$$\lim_{\delta \to 0} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \geq 1. \tag{25}$$

Note that

$$\begin{aligned} &\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \\ = &\min_{j > 1} \inf_{\boldsymbol{\lambda} : \lambda_j > \lambda_1} \sum_{i=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \end{aligned}$$

$$= \min_{j>1} \inf_{\boldsymbol{\lambda}:\lambda_j>\lambda_1} \left(\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_1, \lambda_1\right) + \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_j(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_j, \lambda_j\right) \right)$$
$$= \min_{j>1} \inf_{\lambda\in I} \left(\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_1, \lambda\right) + \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_j(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_j, \lambda\right) \right).$$

Define the function $\Phi(x, y)$ for $x, y \geq 0$ as follows:

$$\Phi(x, y) := \inf_{\lambda\in I} \left(x \cdot \mathrm{kl}\left(\mu_1, \lambda\right) + y \cdot \mathrm{kl}\left(\mu_j, \lambda\right) \right), \tag{26}$$

which frequently appears in the analysis of best arm identification. The properties of the function Φ have been thoroughly explored; see Russo (2020, Lemma 2) for an example. Specifically, the function $\Phi : [0, \infty)^2 \to \mathbb{R}$ has the following properties:

- Φ is continuous on $[0, \infty)^2$.
- For $\mu_1 > \mu_j$, Φ is strictly increasing in the first (resp. second) argument when the second (resp. first) is fixed.
- The infimum in Φ can be replaced with a minimum, and this minimum is attained when $\lambda = \frac{x}{x+y}\mu_1 + \frac{y}{x+y}\mu_j$.

A direct consequence of the last property is that for any finite $y \geq 0$,

$$\lim_{x\to+\infty} \Phi(x, y) = y \cdot \mathrm{kl}\left(\mu_j, \mu_1\right).$$

In the following, we will show $\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)] = \omega(\log(1/\delta))$ via contradiction. Assume, to the contrary, that there exists a finite constant $c > 0$ such that

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]}{\log(1/\delta)} \leq c.$$

Then we have

$$\begin{aligned}
&\liminf_{\delta\to 0} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}\left(\mu_i, \lambda_i\right) \\
&= \liminf_{\delta\to 0} \min_{j>1} \Phi\left(\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]}{\log(1/\delta)}, \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_j(\tau_\delta)]}{\log(1/\delta)} \right) \\
&= \min_{j>1} \liminf_{\delta\to 0} \Phi\left(\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]}{\log(1/\delta)}, \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_j(\tau_\delta)]}{\log(1/\delta)} \right) \\
&\leq \min_{j>1} \Phi\left(\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]}{\log(1/\delta)}, \lim_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_j(\tau_\delta)]}{\log(1/\delta)} \right) \\
&\leq \min_{j>1} \Phi\left(c, \frac{1}{\mathrm{kl}\left(\mu_j, \mu_1\right)} \right) \\
&< \min_{j>1} \lim_{x\to+\infty} \Phi\left(x, \frac{1}{\mathrm{kl}\left(\mu_j, \mu_1\right)} \right)
\end{aligned} \tag{27}$$

$$= \min_{j>1} \frac{\mathrm{kl}(\mu_j, \mu_1)}{\mathrm{kl}(\mu_j, \mu_1)}$$
$$= 1$$

where Line (27) is due to Lemma 10 and the above-stated properties of the function Φ. Note from Equation (24) that for each $j > 1$, $\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_j(\tau_\delta)]}{\log(1/\delta)}$ converges as $\delta \to 0$.

Therefore, we can conclude that

$$\liminf_{\delta \to 0} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)]}{\log(1/\delta)} \cdot \mathrm{kl}(\mu_i, \lambda_i) < 1,$$

which contradicts Inequality (25). Consequently, our claim that $\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)] = \omega(\log(1/\delta))$ holds.

Finally, as $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \geq \mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)]$, it holds that

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \omega(\log(1/\delta))$$

as desired. ■

Appendix D. Proof of Theorem 7

The proof of Theorem 7 consists of three parts. For ease of notation, we write $\gamma = \log\log(1/\delta)$ and omit the dependence of the bandit instance $\boldsymbol{\mu}$ in this appendix. Specifically, we abbreviate $\mathbb{E}_{\boldsymbol{\mu}}$ and $\mathbb{P}_{\boldsymbol{\mu}}$ as $\mathbb{E}$ and $\mathbb{P}$. Furthermore, for each arm $i \in [K]$, we define $\hat{\mu}_{is}$ as the empirical mean of arm i based on its first s pulls. See Table 1 for essential notations used throughout the proofs.

Table 1: Frequently used notations in the proofs.

Symbol	Definition
γ	$\gamma = \log\log(1/\delta)$
ϵ	$\epsilon = \gamma^{-1/4} = \frac{1}{(\log\log(1/\delta))^{1/4}}$
$\beta(\delta)$	$\beta(\delta) = 1 - \min\left\{\frac{1}{\log\log(1/\delta)}, \frac{1}{2}\right\}$
$h(\delta)$	$h(\delta) = \frac{\log(1/\delta)\cdot\gamma}{\epsilon}$
t_r	$t_r = 2^r h(\delta)$ for $r \in \{0\} \cup \mathbb{N}$
$\mathrm{kl}(\mu, \mu')$	$\underline{\mathrm{kl}}(\mu, \mu') = \mathrm{kl}(\mu, \mu') \cdot \mathbb{1}(\mu \leq \mu')$

Proof of correctness Consider any fixed confidence level $\delta \in (0, 1)$. If the algorithm returns a suboptimal arm, then there must exist some $t \in \mathbb{N}$ such that

$$U_1^g(t) < \mu_1 \quad \text{or} \quad \exists i \geq 2, L_i^g(t) > \mu_i.$$

For the best arm (i.e., arm 1), according to Lemma 9, it holds that

$$\mathbb{P}\left(\exists t \in \mathbb{N}, U_1^g(t) < \mu_1\right)$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\exists s \in \mathbb{N}, \hat{\mu}_{1s} \leq \mu_1 \ \text{ and } \mathrm{kl}(\hat{\mu}_{1s}, \mu_1) > \frac{\log(2Ks^2/\delta)}{s}\right) \\
&\leq \sum_{s=1}^{\infty} \mathbb{P}\left(\hat{\mu}_{1s} \leq \mu_1 \ \text{ and } \mathrm{kl}(\hat{\mu}_{1s}, \mu_1) > \frac{\log(2Ks^2/\delta)}{s}\right) \\
&\leq \sum_{s=1}^{\infty} \exp\left(-s \cdot \frac{\log(2Ks^2/\delta)}{s}\right) \\
&= \sum_{s=1}^{\infty} \frac{\delta}{2Ks^2} \\
&\leq \frac{\delta}{K}.
\end{aligned}
$$

Similarly, for any suboptimal arm $i > 1$, we can obtain

$$\mathbb{P}\left(\exists t \in \mathbb{N}, L_i^g(t) > \mu_i\right) \leq \frac{\delta}{K}.$$

Therefore, by applying union bound, we can bound the interested error probability as follows:

$$
\begin{aligned}
\mathbb{P}(i_{\text{out}} \neq 1) &\leq \mathbb{P}\left(\exists t \in \mathbb{N} : \{U_1^g(t) < \mu_1\} \vee \{\exists i \geq 2, L_i^g(t) > \mu_i\}\right) \\
&\leq \mathbb{P}\left(\exists t \in \mathbb{N}, U_1^g(t) < \mu_1\right) + \sum_{i=2}^{K} \mathbb{P}\left(\exists t \in \mathbb{N}, L_i^g(t) > \mu_i\right) \\
&\leq \delta.
\end{aligned}
$$

■

Proof of cumulative regret Let $\epsilon > 0$ and

$$h(\delta) = \frac{\log(1/\delta) \cdot \gamma}{\epsilon}.$$

For the asymptotic analysis, we can pick

$$\epsilon = \gamma^{-1/4} = \frac{1}{(\log\log(1/\delta))^{1/4}}.$$

For $r \in \{0\} \cup \mathbb{N}$, we set $t_r = 2^r h(\delta)$ and define the event

$$\mathcal{E}(r) = \{\text{Algorithm 1 returns after } t_r\}.$$

Therefore,

$$t_0 = h(\delta) = \frac{\log(1/\delta) \cdot \gamma}{\epsilon} \tag{28}$$

Using the linearity of expectation, we can decompose the expected cumulative regret as follows:

$$\begin{aligned}\mathbb{E}[R(\tau_\delta)] &= \mathbb{E}[R(t_0)] + \sum_{r=0}^{\infty} \mathbb{E}\left[(R(t_{r+1}) - R(t_r))\,\mathbb{1}\left[\mathcal{E}(r)\right]\right] \\ &\le \mathbb{E}[R(t_0)] + \sum_{r=0}^{\infty} t_r \Delta_{\max} \mathbb{P}(\mathcal{E}(r)),\end{aligned}$$

where we denote $\Delta_{\max} = \max_{i\in[K]} \Delta_i$.

We introduce one shorthand notation:

$$\Pi^* = \sum_{r=0}^{\infty} t_r \Delta_{\max} \mathbb{P}(\mathcal{E}(r)).$$

In the following steps, we will bound $\mathbb{E}[R(t_0)]$ and Π^* separately. Specifically, we will show $\mathbb{E}[R(t_0)] \le \mathrm{I}^*(\boldsymbol{\mu}) \log(1/\delta) + o(\log(1/\delta))$ and $\Pi^* = o(\log(1/\delta))$.

Bounding $\mathbb{E}[R(t_0)]$. For any $\mu, \mu' \in I$, define $\underline{\mathrm{kl}}(\mu, \mu') = \mathrm{kl}(\mu, \mu') \cdot \mathbb{1}(\mu \le \mu')$. We introduce the following lemmas, which are generalizations of (Lattimore and Szepesvári, 2020, Lemmas 10.7 and 10.8) from Bernoulli distributions to distributions in the exponential family.

Lemma 11 *Let $X_1, X_2, \cdots, X_{t_0}$ be i.i.d. random variables from a one-parameter exponential family with mean μ_1. Further, let $\epsilon > 0$, and define*

$$\psi = \min\left\{ t : \max_{s\in[t_0]} \underline{\mathrm{kl}}(\hat{\mu}_{1s}, \mu_1 - \epsilon) - \frac{f(t)}{s} \le 0 \right\}.$$

Then,

$$\mathbb{E}[\psi] \le 1 + \frac{3V}{\epsilon^2}.$$

Lemma 12 *Let $X_1, X_2, \cdots, X_{t_0}$ be i.i.d. exponential family random variables with mean μ_i. Further, let $\epsilon > 0$, and define*

$$\kappa_i = \sum_{s=1}^{t_0} \mathbb{1}\left\{ \mathrm{kl}(\hat{\mu}_{is}, \mu_1 - \epsilon) \le \frac{f(t_0)}{s} \right\}.$$

Then,

$$\mathbb{E}[\kappa_i] \le \inf_{\epsilon' \in (0, \Delta_i - \epsilon)} \left(\frac{f(t_0)}{\mathrm{kl}(\mu_i + \epsilon', \mu_1 - \epsilon)} + \frac{2V}{\epsilon'^2} \right).$$

Since $f(t) = 3\log t$, it is straightforward to verify $\mathbb{E}[\psi] = o(\log(1/\delta))$ and $\mathbb{E}[\kappa_i] = o(\log(1/\delta))$.

According to the definitions of ψ and κ_i, we can bound $\mathbb{E}[R(t_0)]$ as follows:

$$\mathbb{E}[R(t_0)] = \sum_{i>1} \sum_{t=1}^{t_0} \Delta_i \mathbb{E}\left[\mathbb{1}\{A_t = i\}\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}[\psi]\Delta_{\max} + \sum_{i>1}\Delta_i \mathbb{E}\Big[\sum_{t=\psi+1}^{t_0} \mathbb{1}\{A_t = i\}\Big] \\
&\leq \mathbb{E}[\psi]\Delta_{\max} + \sum_{i>1}\Delta_i \mathbb{E}\Big[\sum_{t=\psi+1}^{t_0} \mathbb{1}\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) \leq \frac{f(t_0)}{N_i(t-1)}\right\}\Big] \\
&\qquad + \sum_{i>1}\Delta_i \mathbb{E}\Big[\sum_{t=\psi+1}^{t_0} \mathbb{1}\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}\right\}\Big] \\
&\leq \mathbb{E}[\psi]\Delta_{\max} + \sum_{i>1}\Delta_i \mathbb{E}[\kappa_i] \\
&\qquad + \sum_{i>1}\Delta_i \mathbb{E}\Big[\sum_{t=\psi+1}^{t_0} \mathbb{1}\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}\right\}\Big].
\end{aligned}
$$

Let

$$
\Pi_1 = \sum_{i>1}\Delta_i \sum_{t=\psi+1}^{t_0} \mathbb{1}\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}\right\}
$$

Then we have

$$
\mathbb{E}[R(t_0)] \leq \mathbb{E}[\psi]\Delta_{\max} + \sum_{i>1}\Delta_i \mathbb{E}[\kappa_i] + \mathbb{E}[\Pi_1]. \tag{29}
$$

Consider

$$
\mathcal{E}_{\exp}(t) = \bigcup_{i\in[K]\setminus\{1\}} \left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}\right\}.
$$

Note that for any $t > \psi$, $U_1^f(t) \geq \mu_1 - \epsilon$. Furthermore, if for some $i \in [K] \setminus \{1\}$, the condition $\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}\right\}$ holds, then $U_i^f(t) < \mu_1 - \epsilon$, implying $A_t^f \neq i$. Therefore, $\mathcal{E}_{\exp}(t)$ is true only if $A_t = A_t^g$.

The following lemma demonstrates that for

$$
T_0 = \min\left\{\psi + \sum_{i>1}\sum_{t=\psi+1}^{t_0} \mathbb{1}\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) \leq \frac{f(t_0)}{N_i(t-1)}\right\} + \gamma\epsilon\log\frac{1}{\delta}, t_0\right\}
$$

and $t \in (\psi, T_0]$, with high probability, the number of times that $\mathcal{E}_{\exp}(t)$ occurs is less than $2\epsilon\log(1/\delta)$.

Lemma 13 *For sufficiently small δ, it holds that*

$$
\mathbb{P}\Big(\sum_{t=\psi+1}^{T_0} \mathbb{1}\left\{\mathcal{E}_{\exp}(t)\right\} \geq \frac{2\log(1/\delta)}{(\log\log(1/\delta))^{1/4}}\Big) \leq \frac{1}{\log(1/\delta)\cdot(\log\log(1/\delta))^{5/4}}.
$$

Note that

$$\mathbb{E}[\Pi_1] = \mathbb{E}\big[\Pi_1 \cdot \mathbb{1}\{T_0 = t_0\}\big] + \mathbb{E}\big[\Pi_1 \cdot \mathbb{1}\{T_0 < t_0\}\big].$$

Recall that $t_0 = \log(1/\delta)\gamma/\epsilon$, $\gamma = \log\log(1/\delta)$, and $\epsilon = \gamma^{-1/4}$. For the case that $T_0 = t_0$, by Lemma 13,

$$\mathbb{E}\big[\Pi_1 \cdot \mathbb{1}\{T_0 = t_0\}\big] \le \frac{1}{t_0} \cdot \Delta_{\max} \cdot t_0 + \left(1 - \frac{1}{t_0}\right)\Delta_{\max} \cdot 2\epsilon \log(1/\delta) = o\left(\log \frac{1}{\delta}\right).$$

Now, consider the case that $T_0 < t_0$. In this case, we have

$$T_0 = \psi + \sum_{i>1} \sum_{t=\psi+1}^{t_0} \mathbb{1}\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) \le \frac{f(t_0)}{N_i(t-1)}\right\} + \gamma\epsilon \log \frac{1}{\delta}.$$

By Lemma 13, with probability $1 - 1/t_0$,

$$\left(N_1(T_0) - N_1(\psi)\right) + 2\epsilon \log \frac{1}{\delta} > \gamma\epsilon \log \frac{1}{\delta}.$$

Therefore, for sufficiently small δ,

$$\mathbb{P}\left(N_1(T_0) > \frac{1}{2}\gamma\epsilon \log \frac{1}{\delta}\right) \ge 1 - \frac{1}{t_0}.$$

Then, from Lemma 9, for sufficiently small δ,

$$\mathbb{P}\left(\forall s > \frac{1}{2}\gamma\epsilon \log \frac{1}{\delta}, \hat{\mu}_{1s} \ge \mu_1 - \frac{\epsilon}{2}\right) \ge 1 - \exp\left(-\frac{\log \delta^{-1} \cdot \gamma\epsilon^3}{16V}\right) \ge 1 - \frac{1}{t_0}.$$

Moreover, for sufficiently small δ and $t \in [t_0]$, if $\hat{\mu}_1(t-1) \ge \mu_1 - \frac{\epsilon}{2}$ and $N_1(t-1) \ge \frac{1}{2}\gamma\epsilon \log \frac{1}{\delta}$, then

$$\begin{aligned}
& L_1^g(t) \ge \mu_1 - \epsilon \\
\Leftarrow\ & \mathrm{kl}\left(\hat{\mu}_1(t-1), \mu_1 - \epsilon\right) > \frac{g(\delta, t_0)}{N_1(t-1)} \\
\Leftarrow\ & \mathrm{kl}\left(\mu_1 - \epsilon/2, \mu_1 - \epsilon\right) > \frac{g(\delta, t_0)}{N_1(t-1)} \\
\Leftarrow\ & \frac{\epsilon^2}{8V} > \frac{\log(2Kt_0^2/\delta)}{\gamma\epsilon/(2) \cdot \log(1/\delta)}. \qquad \text{(which is true for sufficiently small } \delta\text{)}
\end{aligned}$$

Hence, we have

$$\mathbb{P}\left(\forall t \in [T_0, t_0], L_1^g(t) \ge \mu_1 - \epsilon\right) \ge 1 - \frac{2}{t_0}. \tag{30}$$

The following lemma shows that each suboptimal arm is pulled within the optimal range with high probability.

Lemma 14 *Assuming that $L_1^g(t) \geq \mu_1 - \epsilon$ holds for all $t \geq T_0$, for sufficiently small δ,*

$$\mathbb{P}\bigg(\sum_{t=T_0+1}^{t_0} \mathbb{1}\left\{ A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)} \right\} > \max\left\{ \frac{g(\delta, t_0)}{\mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon)}, \frac{2V\log(t_0)}{\epsilon^2} \right\} \bigg) \leq \frac{1}{t_0}.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\big[\Pi_1 \cdot \mathbb{1}\{T_0 < t_0\}\big] \\
&= \sum_{i>1} \Delta_i \mathbb{E}\bigg[\sum_{t=\psi+1}^{t_0} \mathbb{1}\left\{ T_0 < t_0, A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)} \right\}\bigg] \\
&\leq \sum_{i>1} \Delta_i \mathbb{E}\bigg[\sum_{t=\psi+1}^{T_0} \mathbb{1}\left\{ A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)} \right\}\bigg] \\
&\quad + \sum_{i>1} \Delta_i \mathbb{E}\bigg[\sum_{t=T_0+1}^{t_0} \mathbb{1}\left\{ A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}, \forall t \in [T_0, t_0], L_1^g(t) \geq \mu_1 - \epsilon \right\}\bigg] \\
&\quad + t_0 \Delta_{\max} \mathbb{P}\Big(\exists t \in [T_0, t_0] : L_1^g(t) < \mu_1 - \epsilon \Big) \\
&\leq \Delta_{\max} 2\epsilon \log(1/\delta) + t_0 \Delta_{\max} \cdot \mathbb{P}\bigg(\sum_{t=\psi+1}^{T_0} \mathbb{1}\left\{ \mathcal{E}_{\exp}(t) \right\} \geq 2\epsilon \log(1/\delta) \bigg) \qquad \text{(due to Lemma 13)} \\
&\quad + \sum_{i>1} \Delta_i \cdot t_0 \cdot \frac{1}{t_0} + \sum_{i>1} \Delta_i \max\left\{ \frac{g(\delta, t_0)}{\mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon)}, \frac{2V\log(t_0)}{\epsilon^2} \right\} \qquad \text{(due to Lemma 14)} \\
&\quad + t_0 \cdot \Delta_{\max} \cdot \frac{2}{t_0} \qquad \text{(due to Inequality (30))} \\
&\leq \Delta_{\max} 2\epsilon \log(1/\delta) + 3\Delta_{\max} + \sum_{i>1} \Delta_i + \sum_{i>1} \Delta_i \max\left\{ \frac{g(\delta, t_0)}{\mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon)}, \frac{2V\log(t_0)}{\epsilon^2} \right\} \\
&\leq o\bigg(\log \frac{1}{\delta} \bigg) + \sum_{i>1} \frac{\Delta_i \log(1/\delta)}{\mathrm{kl}(\mu_i, \mu_1)},
\end{aligned}$$

where in the last inequality, we use the fact that $g(\delta, t) = \log(2Kt^2/\delta)$.

Altogether, we can obtain

$$\mathbb{E}[\Pi_1] \leq \sum_{i>1} \frac{\Delta_i \log(1/\delta)}{\mathrm{kl}(\mu_i, \mu_1)} + o\bigg(\log \frac{1}{\delta} \bigg).$$

Combining with (29), we arrive at

$$\begin{aligned}
\mathbb{E}[R(t_0)] &\leq \sum_{i>1} \frac{\Delta_i \log(1/\delta)}{\mathrm{kl}(\mu_i, \mu_1)} + o\bigg(\log \frac{1}{\delta} \bigg) \\
&= \mathrm{I}^*(\boldsymbol{\mu}) \log(1/\delta) + o(\log(1/\delta))
\end{aligned}$$

as desired.

Bounding Π^*. To bound Π^*, it suffices to show the following lemma. The proof of Lemma 15 requires introducing the events $\mathcal{E}_0(r)$, $\mathcal{E}_1(r)$, $\mathcal{E}_2(r)$ and $\mathcal{E}_3(r)$. We refer to Appendix E.1 for details.

Lemma 15 *For sufficiently small δ, it holds that*

$$\mathbb{P}(\mathcal{E}(r)) \leq \frac{14}{t_r^2}.$$

Then we have

$$\begin{aligned}\Pi^* &= \sum_{r=0}^{\infty} t_r \Delta_{\max} \mathbb{P}(\mathcal{E}(r)) \\ &\leq \sum_{r=0}^{\infty} \frac{14\Delta_{\max}}{t_r} \\ &\leq \frac{14\Delta_{\max}}{h(\delta)} \\ &= o(\log(1/\delta)).\end{aligned}$$

The proof of expected cumulative regret is complete. ■

Proof of sample complexity Similar to the proof of expected cumulative regret, we can decompose the sample complexity as follows:

$$\mathbb{E}[\tau_\delta] = t_0 + \sum_{r=0}^{\infty} \mathbb{P}\left(\mathcal{E}(r)\right)(t_{r+1} - t_r).$$

Using Lemma 15, we have

$$\sum_{r=0}^{\infty} \mathbb{P}\left(\mathcal{E}(r)\right)(t_{r+1} - t_r) \leq \sum_{r=0}^{\infty} \frac{14}{t_r^2} \cdot t_r = \sum_{r=0}^{\infty} \frac{14}{2^r h(\delta)} = o\left(\log \frac{1}{\delta}\right).$$

Therefore, we can conclude that

$$\lim_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta) \cdot (\log\log(1/\delta))^2} = \lim_{\delta \to 0} \frac{t_0}{\log(1/\delta) \cdot (\log\log(1/\delta))^2} = 0.$$

■

Appendix E. Proofs of Supporting Lemmas for Theorem 7

Proof of Lemma 11 Recall V is the upper bound of the variances of the exponential family. According to Lemma 9, we have

$$\mathbb{P}(\psi > t) \leq \mathbb{P}\left(\exists 1 \leq s \leq t_0 : \underline{\mathrm{kl}}(\hat{\mu}_{1s}, \mu_1 - \epsilon) > \frac{f(t)}{s}\right)$$

$$
\begin{aligned}
&\leq \sum_{s=1}^{t_0} \mathbb{P}\bigg(\underline{\mathrm{kl}}(\hat{\mu}_{1s}, \mu_1 - \epsilon) > \frac{f(t)}{s} \bigg) \\
&= \sum_{s=1}^{t_0} \mathbb{P}\bigg(\mathrm{kl}(\hat{\mu}_{1s}, \mu_1 - \epsilon) > \frac{f(t)}{s}, \hat{\mu}_{1s} < \mu_1 - \epsilon \bigg) \\
&\leq \sum_{s=1}^{t_0} \mathbb{P}\bigg(\mathrm{kl}(\hat{\mu}_{1s}, \mu_1) > \frac{f(t)}{s} + \frac{\epsilon^2}{2V}, \hat{\mu}_{1s} < \mu_1 \bigg) \\
&\leq \sum_{s=1}^{t_0} \exp\bigg(-s\Big(\frac{\epsilon^2}{2V} + \frac{f(t)}{s} \Big) \bigg) \\
&\leq \frac{1}{\exp(f(t))} \sum_{s=1}^{t_0} \exp(-\frac{s\epsilon^2}{2V}) \\
&\leq \frac{2V}{t^3\epsilon^2}.
\end{aligned}
$$

Therefore, we can obtain

$$
\mathbb{E}[\psi] = 1 + \sum_{t=1}^{\infty} \mathbb{P}(\psi > t) \leq 1 + \frac{3V}{\epsilon^2}.
$$

■

Proof of Lemma 12 Let $\epsilon' \in (0, \Delta_i - \epsilon)$ and $u = \frac{f(t_0)}{\mathrm{kl}(\mu_i + \epsilon', \mu_1 - \epsilon)}$. Then

$$
\begin{aligned}
\mathbb{E}[\kappa_i] &= \sum_{s=1}^{t_0} \mathbb{P}\left(\mathrm{kl}(\hat{\mu}_{is}, \mu_1 - \epsilon) \leq \frac{f(t_0)}{s} \right) \\
&\leq \sum_{s=1}^{t_0} \mathbb{P}\left(\hat{\mu}_{is} \geq \mu_i + \epsilon' \ \text{ or } \ \mathrm{kl}(\mu_i + \epsilon', \mu_1 - \epsilon) \leq \frac{f(t_0)}{s} \right) \\
&\leq u + \sum_{s=\lceil u \rceil}^{t_0} \mathbb{P}(\hat{\mu}_{is} \geq \mu_i + \epsilon') \\
&\leq u + \sum_{s=1}^{\infty} \exp\Big(-s \cdot \mathrm{kl}(\mu_i + \epsilon', \mu_i) \Big) && \text{(due to Lemma 9)} \\
&\leq \frac{f(t_0)}{\mathrm{kl}(\mu_i + \epsilon', \mu_1 - \epsilon)} + \frac{1}{\mathrm{kl}(\mu_i + \epsilon', \mu_i)} \\
&\leq \frac{f(t_0)}{\mathrm{kl}(\mu_i + \epsilon', \mu_1 - \epsilon)} + \frac{2V}{\epsilon'^2}. && \text{(due to Lemma 8)}
\end{aligned}
$$

■

Proof of Lemma 13 Note that the claim of Lemma 13 is equivalent to showing that

$$\mathbb{P}\Big(\sum_{t=\psi+1}^{T_0} \mathbb{1}\left\{\mathcal{E}_{\exp}(t)\right\} \geq 2\log(1/\delta) \Big) \leq \frac{1}{t_0}.$$

We note that for $t > \psi$, one of the following three disjoint events occurs:

- $A_t = 1$;
- for some $i > 1$, $A_t = i$ and $\mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}$;
- for some $i > 1$, $A_t = i$ and $\mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) \leq \frac{f(t_0)}{N_i(t-1)}$.

Besides, for $t > \psi$, the event $\mathcal{E}_{\exp}(t)$ only happens when the coin toss yields tails, and the condition

$$\mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}$$

holds true. Therefore, by the definition of $\mathcal{E}_{\exp}(t)$, $\sum_{t=\psi+1}^{T_0} \mathbb{1}\left\{\mathcal{E}_{\exp}(t)\right\}$ is at most the number of heads that we toss $\gamma\epsilon\log\frac{1}{\delta}$ coins with bias $1-\beta(\delta) \leq 1/\gamma$. From Hoeffding's bound, we have that for sufficiently small δ,

$$\begin{aligned}
\mathbb{P}\Big(\sum_{t=\psi+1}^{T_0} \mathbb{1}\left\{\mathcal{E}_{\exp}(t)\right\}) - \gamma\epsilon\log\frac{1}{\delta}\cdot\frac{1}{\gamma} \geq \epsilon\log\frac{1}{\delta} \Big) &\leq \exp\left(-\frac{2\epsilon^2\big(\log\frac{1}{\delta}\big)^2}{\gamma\epsilon\log\frac{1}{\delta}} \right) \\
&\leq \exp\left(-\frac{2\epsilon\log\frac{1}{\delta}}{\gamma} \right) \\
&\leq \frac{1}{t_0},
\end{aligned}$$

which completes the proof of Lemma 13. ■

Proof of Lemma 14 Note that since $T_0 > \psi$, $U_1^f(t) > \mu_1 - \epsilon$ for $t > T_0$. Therefore, the event $\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}\right\}$ occurs only when $A_t = A_t^g = i$. Besides, the Algorithm returns when

$$L_1^g(t) > \mu_1 - \epsilon \;\text{ and }\; \forall i \in [K]\setminus\{1\}, U_i^g(t) \leq \mu_1 - \epsilon.$$

After pulling arm i for $\max\left\{\frac{g(\delta,t_0)}{\mathrm{kl}(\mu_i+\epsilon,\mu_1-\epsilon)}, \frac{2V\log(t_0)}{\epsilon^2}\right\}$ times, by Lemma 9,

$$\mathbb{P}(\hat{\mu}_i(t-1) \leq \mu_i + \epsilon) \geq 1 - \frac{1}{t_0}.$$

Furthermore, conditioned on the event that $\hat{\mu}_i(t-1) \leq \mu_i + \epsilon$ and arm i is pulled for more than $\max\left\{\frac{g(\delta,t_0)}{\mathrm{kl}(\mu_i+\epsilon,\mu_1-\epsilon)}, \frac{2V\log(t_0)}{\epsilon^2}\right\}$ times, we have

$$U_i^g(t) \leq \sup\left\{\mu \in I : \mathrm{kl}(\hat{\mu}_i(t-1), \mu) \leq \frac{g(\delta,t_0)}{N_i(t-1)}\right\}$$

$$< \sup\left\{\mu \in I : \mathrm{kl}(\hat{\mu}_i(t-1), \mu) \le \mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon)\right\}$$
$$\le \mu_1 - \epsilon.$$

Recall that $g(\delta, t) = \log(2Kt^2/\delta)$ and $f(t) = 3\log t$. For sufficiently small δ, $g(\delta, t) > f(t)$ for all $t \in [t_0]$. Therefore, for all $t \in [t_0]$, $U_i^f(t) < U_i^g(t) < \mu_1 - \epsilon$.

According to Algorithm 1, if it happens that $A_t^g = i \neq 1$, then

1. $U_1^f(t) > L_1^g(t) > \mu_1 - \epsilon > U_i^g(t) > \max_{j\in[K]\setminus\{1,i\}} U_j^g(t) \ge \max_{j\in[K]\setminus\{1,i\}} U_j^f(t)$, which implies that $A_t^f = 1$.
2. $L_1^g(t) = L_{A_t^f}^g(t) \ge \mu_1 - \epsilon \ge \max_{j\in[K]\setminus\{1\}} U_j^g(t)$, which means that the algorithm returns.

Therefore, it cannot be the case that $A_t^g = i \neq 1$, and we obtain:

$$\mathbb{P}\Bigg(\sum_{t=T_0+1}^{t_0} \mathbb{1}\left\{A_t = i, \mathrm{kl}(\hat{\mu}_i(t-1), \mu_1 - \epsilon) > \frac{f(t_0)}{N_i(t-1)}\right\}$$
$$> \max\left\{\frac{g(\delta, t_0)}{\mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon)}, \frac{2V\log(t_0)}{\epsilon^2}\right\}\Bigg) \le \frac{1}{t_0}.$$

■

E.1 Proof of Lemma 15

Let L_r be the number of heads in coin tosses before $t_r/2$, and

$$M(r) = \frac{2V\log(Kt_r^2)}{\epsilon^2} + \frac{\max\{f(t_r), g(\delta, t_r)\}}{\min_{i>1}\mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon)}.$$

We define the following events:

$$\mathcal{E}_0(r) = \left\{L_r > \frac{t_r}{8}\right\},$$
$$\mathcal{E}_1(r) = \left\{\forall t \ge \frac{t_r}{16}, U_1^f(t) \ge \mu_1 - \epsilon\right\},$$
$$\mathcal{E}_2(r) = \left\{\forall i \in [K]\setminus\{1\}, t \le t_r \text{ and } N_i(t) > M(r) : U_i^f(t) < \mu_1 - \epsilon, U_i^g(t) \le \mu_1 - \epsilon\right\},$$
$$\mathcal{E}_3(r) = \left\{\forall t \in (t_r/2, t_r], L_1^g(t) \ge \mu_1 - \epsilon\right\}.$$

To prove Lemma 15, we require the following lemmas, whose proofs are deferred to the end of this subsection.

Lemma 16 *For sufficiently small δ,*

$$\mathbb{P}(\mathcal{E}_0(r)) \ge 1 - \frac{1}{t_r^2}.$$

Lemma 17 *For sufficiently small δ,*

$$\mathbb{P}(\mathcal{E}_1(r)) \geq 1 - \frac{1}{t_r^2}.$$

Lemma 18 *For sufficiently small δ,*

$$\mathbb{P}(\mathcal{E}_2(r)) \geq 1 - \frac{1}{t_r^2}.$$

Lemma 19 *If $\mathcal{E}_0(r)$, $\mathcal{E}_1(r)$ and $\mathcal{E}_2(r)$ are true, the number of pulls of the optimal arm by time $t_r/2$ is at least*

$$\frac{t_r}{16} - KM(r).$$

Besides, for sufficiently small δ,

$$\mathbb{P}\left(\mathcal{E}_3(r) \mid \mathcal{E}_0(r), \mathcal{E}_1(r), \mathcal{E}_2(r)\right) \geq 1 - \frac{1}{t_r^2}.$$

Proof of Lemma 15 Assume that $\mathcal{E}_0(r)$, $\mathcal{E}_1(r)$, $\mathcal{E}_2(r)$ and $\mathcal{E}_3(r)$ are true. We first divide the time steps in $t \in [t_r]$ into three disjoint sets. We let

$$\mathcal{T}_1(r) = \left\{ t \in (t_r/2, t_r] \;\middle|\; \max_{i \in [K]\setminus\{1\}} U_i^g(t) < \mu_1 - \epsilon \text{ and } \max_{i \in [K]\setminus\{1\}} U_i^f(t) < U_1^f(t) \right\}$$

$$\mathcal{T}_2^g(r) = \left\{ t \in (t_r/2, t_r] \;\middle|\; \max_{i \in [K]\setminus\{1\}} U_i^g(t) \geq \mu_1 - \epsilon \text{ and } \max_{i \in [K]\setminus\{1\}} U_i^f(t) < U_1^f(t) \right\}.$$

and

$$\mathcal{T}_2^f(r) = \left\{ t \in (t_r/2, t_r] \;\middle|\; \max_{i \in [K]\setminus\{1\}} U_i^f(t) \geq U_1^f(t) \right\}.$$

Let Y_t be the indicator variable such that

$$Y_t = \begin{cases} 1 & \text{if the coin toss at time } t \text{ results in heads,} \\ 0 & \text{if the coin toss at time } t \text{ results in tails.} \end{cases}$$

Assume $A_t = A_t^g \in [K] \setminus \{1\}$ occurs for at least $(K-1)M(r)$ times by time $\hat{t}$. Since $\mathcal{E}_2(r)$ is true, for $t \in (\hat{t}, t_r]$,

$$\max_{i \in [K]\setminus\{1\}} U_i^g(t) < \mu_1 - \epsilon.$$

Therefore,

$$\sum_{t \in \mathcal{T}_2^g(r)} \mathbb{1}\left\{A_t^g \in [K] \setminus \{1\}, Y_t = 0\right\} \leq (K-1)M(r).$$

Note that

$$\mathbb{E}\bigg[\mathbb{1}\{A_t^g \in [K]\setminus\{1\}, Y_t = 0\} \;\bigg|\; \max_{i\in[K]\setminus\{1\}} U_i^f(t) < U_1^f(t)\bigg] = \mathbb{P}(Y_t = 0) = 1-\beta(\delta).$$

Consider the independent Bernoulli random variable $\{Z_i\}_{i\geq 1}$ with bias $1-\beta(\delta)$. Then we have that for fixed L,

$$\mathbb{P}\bigg(|\mathcal{T}_2^g(r)| \geq L\bigg) \leq \mathbb{P}\bigg(\sum_{i=1}^{L} Z_i < (K-1)M(r)\bigg).$$

Let

$$E_g = \bigg\{\sum_{i=1}^{t_r/8} Z_i \geq \frac{t_r}{16}(1-\beta(\delta))\bigg\}.$$

Applying Hoeffding's bound, we have

$$\begin{aligned}\mathbb{P}\big(E_g^c\big) &\leq \exp\left(-\frac{2\frac{t_r^2}{16^2}(1-\beta(\delta))^2}{t_r/8}\right)\\ &\lesssim \frac{1}{t_r^2}. \qquad \text{(for sufficiently small } \delta)\end{aligned}$$

where we use $\lesssim$ as shorthand for Big O notation, indicating asymptotic behavior.

Note that for sufficiently small δ,

$$\frac{t_r}{16}(1-\beta(\delta)) \geq (K-1)M(r)$$

Therefore, setting $L = t_r/8$, we can obtain

$$\mathbb{P}\bigg(|\mathcal{T}_2^g(r)| < \frac{t_r}{8}\bigg) \geq \mathbb{P}(\mathcal{E}_0(r), \mathcal{E}_1(r), \mathcal{E}_2(r), \mathcal{E}_3(r), E_g) \geq 1 - \frac{5}{t_r^2}.$$

Similarly, assume $A_t = A_t^f \in [K]\setminus\{1\}$ occurs for at least $(K-1)M(r)$ times by time $\hat{t}$. Since $\mathcal{E}_1(r)$ and $\mathcal{E}_2(r)$ are true, for $t \in (\hat{t}, t_r]$,

$$\max_{i\in[K]\setminus\{1\}} U_i^f(t) < \mu_1 - \epsilon \leq U_1^f(t).$$

Therefore,

$$\sum_{t\in\mathcal{T}_2^f(r)} \mathbb{1}\left\{A_t^f \in [K]\setminus\{1\}, Y_t = 1\right\} \leq (K-1)M(r).$$

Note that

$$\mathbb{P}\bigg(A_t = A_t^f \;\bigg|\; \max_{i>1} U_i^f(t) \geq U_1^f\bigg) = \beta(\delta).$$

Consider the collection of independent Bernoulli random variables $\{X_i\}_{i\geq 1}$ with bias $\beta(\delta)$. Then we have that for fixed L,

$$\mathbb{P}\bigg(|\mathcal{T}_2^f(r)| \geq L\bigg) \leq \mathbb{P}\bigg(\sum_{i=1}^{L} X_i < (K-1)M(r)\bigg).$$

Let

$$E_f = \bigg\{\sum_{i=1}^{t_r/8} X_i \geq \frac{t_r}{16}\beta(\delta)\bigg\}.$$

Similarly, for sufficiently small δ, we have $\mathbb{P}(E_f^c) \leq 1/t_r^2$ and

$$\mathbb{P}\bigg(|\mathcal{T}_2^f(r)| < \frac{t_r}{8}\bigg) \geq \mathbb{P}(\mathcal{E}_0(r), \mathcal{E}_1(r), \mathcal{E}_2(r), \mathcal{E}_3(r), E_f) \geq 1 - \frac{5}{t_r^2}.$$

Now, we have with probability at least $1 - 14/t_r^2$, it holds that

$$\mathcal{E}_0(r), \mathcal{E}_1(r), \mathcal{E}_2(r), \mathcal{E}_3(r), |\mathcal{T}_2^f(r)| < \frac{t_r}{8}, |\mathcal{T}_2^g(r)| < \frac{t_r}{8}.$$

Therefore, with probability at least $1 - 14/t_r^2$,

$$|\mathcal{T}_1(r)| \geq t_r/2 - t_r/8 - t_r/8 > 0.$$

Then for any $t \in \mathcal{T}_1(r)$, we have

1. $A_t^f = 1$ because $U_1^f(t) > \max_{i\in[K]\setminus\{1\}} U_i^f(t)$;
2. $L_1^g(t) \geq \mu_1 - \epsilon$ because $\mathcal{E}_3(r)$ is true;
3. $\max_{i\in[K]\setminus\{1\}} U_i^g(t) < \mu_1 - \epsilon$.

Hence,

$$L_{A_t^f}^g(t) = L_1^g(t) \geq \mu_1 - \epsilon \geq \max_{i\in[K]\setminus\{1\}} U_i^g(t),$$

which indicates that the algorithm returns.

This completes the proof of Lemma 15. ■

Proof of Lemma 16 Since

$$\beta(\delta) = 1 - \min\left\{\frac{1}{\gamma}, \frac{1}{2}\right\} \geq \frac{1}{2},$$

by Hoeffding's bound, we have

$$\mathbb{P}\bigg(\mathcal{E}_0(r)^c\bigg) = \mathbb{P}\bigg(L_r \leq \frac{t_r}{8}\bigg) \leq \mathbb{P}\bigg(L_r - \beta(\delta)\cdot\frac{t_r}{2} < -\frac{1}{8}t_r\bigg)$$

$$\leq \exp\bigg(-\frac{t_r}{32}\bigg)$$
$$\lesssim \frac{1}{t_r^2}$$

for sufficiently small δ.

■

Proof of Lemma 17

$$\begin{aligned}
\mathbb{P}\bigg(\mathcal{E}_1(r)^c\bigg) &= \mathbb{P}\bigg(\exists t > \frac{t_r}{16}, \hat{\mu}_1(t-1) < \mu_1 - \epsilon, \mathrm{kl}(\hat{\mu}_1(t-1), \mu_1 - \epsilon) \geq \frac{f(t)}{N_i(t-1)}\bigg) \\
&\leq \mathbb{P}\bigg(\exists s \geq 1, \hat{\mu}_{1s} < \mu_1 - \epsilon, \mathrm{kl}(\hat{\mu}_{1s}, \mu_1 - \epsilon) \geq \frac{f(t_r/16)}{s}\bigg) \\
&\leq \mathbb{P}\bigg(\exists s \geq 1, \hat{\mu}_{1s} < \mu_1 - \epsilon, \mathrm{kl}(\hat{\mu}_{1s}, \mu_1) \geq \frac{f(t_r/16)}{s} + \frac{\epsilon^2}{2V}\bigg) \quad \text{(due to Lemma 8)} \\
&\leq \sum_{s=1}^{\infty} \exp\bigg(-s\bigg(\frac{\epsilon^2}{2V} + \frac{f(t_r/16)}{s}\bigg)\bigg) \quad \text{(due to Lemma 9)} \\
&\leq \frac{1}{e^{f(t_r/16)}} \sum_{s=1}^{\infty} \exp\big(-s\epsilon^2/(2V)\big) \\
&\leq \frac{2V}{\epsilon^2 \cdot e^{f(t_r/16)}} \\
&\lesssim \frac{1}{t_r^2}.
\end{aligned}$$

■

Proof of Lemma 18 From Lemma 9, we have

$$\begin{aligned}
\mathbb{P}(\exists i \in [K] \setminus \{1\}, \exists s > M(r), \hat{\mu}_{is} > \mu_i + \epsilon) &\leq \sum_{i \in [K]\setminus\{1\}} \mathbb{P}(\exists s > M(r), \hat{\mu}_{is} > \mu_i + \epsilon) \\
&\leq K \exp\bigg(-\frac{M(r)\epsilon^2}{2V}\bigg) \\
&\leq \frac{1}{t_r^2}.
\end{aligned}$$

Then, with probability at least $1 - \frac{1}{t_r^2}$, we have for all $i > 1$ and $s > M(r)$,

$$\mathrm{kl}(\hat{\mu}_{is}, \mu_1 - \epsilon) \geq \mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon) > \frac{f(t_r)}{s}; \qquad \text{(which implies } U_i^f(t) < \mu_1 - \epsilon)$$
$$\mathrm{kl}(\hat{\mu}_{is}, \mu_1 - \epsilon) \geq \mathrm{kl}(\mu_i + \epsilon, \mu_1 - \epsilon) \geq \frac{g(\delta, t_r)}{s}. \qquad \text{(which implies } U_i^g(t) \leq \mu_1 - \epsilon)$$

■

Proof of Lemma 19 Since $\mathcal{E}_0(r)$ and $\mathcal{E}_1(r)$ are true, $L_r \geq \frac{t_r}{8}$ and when $t > \frac{t_r}{16}$, $U_1(t) > \mu_1 - \epsilon$. Therefore, among the L_r time steps where the result of the coin toss is heads, there are at least $L_r - \frac{t_r}{16} \geq \frac{t_r}{16}$ time steps where if $U_i(t) < \mu_1 - \epsilon$ for all $i > 1$, then $A_t = 1$.

Since $\mathcal{E}_2(r)$ is true, after at most $KM(r)$ pulls on the suboptimal arms, we have $U_i(t) < \mu_1 - \epsilon$ for all $i > 1$. Therefore, if $\mathcal{E}_0(r), \mathcal{E}_1(r)$, and $\mathcal{E}_2(r)$ are true, the number of pulls of the optimal arm by time $t_r/2$ is at least $t_r/16 - KM(r)$, which completes the proof of the first statement.

For the second statement, consider any $t \in (t_r/2, t_r]$. We will show by contradiction that if $\hat{\mu}_1(t-1) > \mu_1 - \epsilon/2$, then $L_1^g(t) \geq \mu_1 - \epsilon$.

Suppose that $L_1^g(t) < \mu_1 - \epsilon$. Then we have

$$\begin{aligned}\frac{g(\delta, t_r)}{N_1(t-1)} &\geq \frac{g(\delta, t)}{N_1(t-1)} \\ &\geq \mathrm{kl}(\hat{\mu}_1(t-1), \mu_1 - \epsilon) \\ &\geq \mathrm{kl}(\mu_1 - \epsilon/2, \mu_1 - \epsilon).\end{aligned}$$

Therefore, together with Lemma 8, we can obtain

$$N_1(t-1) \leq \frac{g(\delta, t_r)}{\mathrm{kl}(\mu_1 - \epsilon/2, \mu_1 - \epsilon)} \leq \frac{8V g(\delta, t_r)}{\epsilon^2}.$$

However, according to the first statement,

$$N_1(t-1) \geq \frac{t_r}{16} - KM(r).$$

Note that for sufficiently small δ,

$$\frac{t_r}{16} - KM(r) \gtrsim \frac{8V g(\delta, t_r)}{\epsilon^2},$$

which leads to a contradiction. Therefore, we can establish that $L_1^g(t) \geq \mu_1 - \epsilon$.

Furthermore, from Lemma 9,

$$\mathbb{P}(\exists t \in (t_r/2, t_r], \hat{\mu}_1(t-1) < \mu_1 - \epsilon/2) \leq \exp\left(\frac{-(t_r/16 - KM(r)) \cdot \epsilon^2}{8V}\right) \lesssim \frac{1}{t_r^2}.$$

Altogether, we have for sufficiently small δ,

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_3(r) \mid \mathcal{E}_0(r), \mathcal{E}_1(r), \mathcal{E}_2(r)\right) &\geq \mathbb{P}(\forall t \in (t_r/2, t_r], \hat{\mu}_1(t-1) \geq \mu_1 - \epsilon/2) \\ &\geq 1 - \frac{1}{t_r^2}.\end{aligned}$$

■

Appendix F. Proof of the Uniformity in the Allocation in Example 1

Consider a two-armed Bernoulli bandit instance $\boldsymbol{\mu} = (1-\mu, \mu)$ with $\mu \in (0, 1/2)$. For any δ-PAC BAI algorithm, it holds that

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq \Gamma^*(\boldsymbol{\mu})$$

where

$$\begin{aligned}
\Gamma^*(\boldsymbol{\mu})^{-1} :&= \sup_{w \in \mathcal{P}_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{i=1}^K w_i \mathrm{kl}\left(\mu_i, \lambda_i\right) \right) \\
&= \sup_{w \in \mathcal{P}_K} \inf_{\lambda_2 > \lambda_1} \left(w_1 \mathrm{kl}\left(\mu_1, \lambda_1\right) + w_2 \mathrm{kl}\left(\mu_2, \lambda_2\right) \right) \\
&= \sup_{w \in \mathcal{P}_K} \inf_{\mu_1 < \lambda < \mu_2} \left(w_1 \mathrm{kl}\left(\mu_1, \lambda\right) + w_2 \mathrm{kl}\left(\mu_2, \lambda\right) \right) \\
&= \sup_{w \in \mathcal{P}_K} \inf_{1-\mu < \lambda < \mu} \left(w_1 \mathrm{kl}\left(1-\mu, \lambda\right) + w_2 \mathrm{kl}\left(\mu, \lambda\right) \right).
\end{aligned}$$

Substituting the KL divergence for Bernoulli distributions yields

$$\begin{aligned}
\Gamma^*(\boldsymbol{\mu})^{-1} &= \sup_{w \in \mathcal{P}_K} \inf_{1-\mu < \lambda < \mu} \Bigg(w_1 \left((1-\mu) \log\left(\frac{1-\mu}{\lambda}\right) + \mu \log\left(\frac{\mu}{1-\lambda}\right) \right) \\
&\qquad\qquad + w_2 \left(\mu \log\left(\frac{\mu}{\lambda}\right) + (1-\mu) \log\left(\frac{1-\mu}{1-\lambda}\right) \right) \Bigg) \\
&= \sup_{w \in \mathcal{P}_K} \inf_{1-\mu < \lambda < \mu} \Big(- (w_1(1-\mu) + w_2 \mu) \log \lambda - (w_1 \mu + w_2(1-\mu)) \log(1-\lambda) \\
&\qquad\qquad + (\mu \log \mu + (1-\mu) \log(1-\mu)) \Big)
\end{aligned}$$

Taking the derivative with respect to λ reveals that the inner infimum is attained at $\lambda_* := w_1(1-\mu) + w_2 \mu$. Consequently,

$$\Gamma^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \mathcal{P}_K} \Big(- \lambda_* \log \lambda_* - (1-\lambda_*) \log(1-\lambda_*) + (\mu \log \mu + (1-\mu) \log(1-\mu)) \Big).$$

It is readily seen that the outer supremum is maximized when $\lambda_* = 1 - \lambda_*$, which implies $w_1 = w_2 = 1/2$. Therefore, the optimal allocation is uniform for **all** $\mu \in (0, 1/2)$. (The special case that $\mu = 1/2$ is excluded from our analysis as the optimal arm is no longer unique.)

References

Rajeev Agrawal. Sample mean based index policies by $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory (COLT)*, pages 41–53, 2010.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

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

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

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

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

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

Rémy Degenne. On the existence of a complexity in fixed budget bandit identification. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1131–1154. PMLR, 2023.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019a.

Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1988–1996. PMLR, 2019b.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, pages 255–270, 2002.

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

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

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

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

Tianyuan Jin, Yu Yang, Jing Tang, Xiaokui Xiao, and Pan Xu. Optimal batched best arm identification. *Advances in Neural Information Processing Systems*, 37:134947–134980, 2024.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

Kellen Kanarios, Qining Zhang, and Lei Ying. Cost aware best arm identification. In *Reinforcement Learning Conference (RLC)*, 2024.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

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

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

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

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, 2017.

Arpan Mukherjee and Ali Tajer. Best arm identification in stochastic bandits: Beyond β-optimality. *arXiv preprint arXiv:2301.03785*, 2023.

Chao Qin and Daniel Russo. Optimizing adaptive experiments: A unified approach to regret minimization and best-arm identification. *arXiv preprint arXiv:2402.10592*, 2024.

Daniel Russo. Simple Bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

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

Qining Zhang and Lei Ying. Fast and regret optimal best arm identification: Fundamental limits and low-complexity algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yao Zhao, Connor Stephens, Csaba Szepesvári, and Kwang-Sung Jun. Revisiting simple regret: Fast rates for returning a good arm. In *International Conference on Machine Learning*, pages 42110–42158. PMLR, 2023.

Zixin Zhong, Wang Chi Cheung, and Vincent Tan. Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *Transactions on Machine Learning Research*, 2023.