On Minimizing Regret in Fixed-Confidence ε-Best Arm Identification

Tianyuan Jin*
Data Science and Analytics Thrust
The Hong Kong University of Science and Technology (Guangzhou)
tianyuanjin@hkust-gz.edu.cn

Junwen Yang
Institute of Operations Research and Analytics
National University of Singapore
junwen_yang@u.nus.edu

Vincent Y. F. Tan
Department of Mathematics
Department of Electrical and Computer Engineering
Institute of Operations Research and Analytics
National University of Singapore
vtan@nus.edu.sg

Abstract

This paper studies the ε-best arm identification problem (ε-BAI) for K-armed bandits under the fixed-confidence setting. Departing from the usual objective of identifying the best arm with minimal sample complexity, we focus on minimizing the cumulative regret of identifying an ε-best arm. We consider the usual asymptotic regime in which the error probability δ of identifying an ε-best arm vanishes. In this setting, we show a novel instance-dependent lower bound. In particular, we show that any family of (ε, δ)-PAC algorithms must incur an expected cumulative regret up to the stopping time that scales as $\log(1/\delta)$, with an instance-dependent constant $C(\boldsymbol{\mu})$. This is proved using a regret-aware change-of-measure technique that allocates information given a budget on the regret. Motivated by this lower bound, we devise an algorithm, `KLUCB+Greedy`, that achieves the same scaling in δ, i.e., $\log(1/\delta)$, but with a different constant, which we call $C^*(\boldsymbol{\mu})$. `KLUCB+Greedy` has two main benefits: (i) it is computationally efficient as $C^*(\boldsymbol{\mu})$ admits a closed-form expression; (ii) in broad classes of reward distributions such as Gaussian distributions, $C^*(\boldsymbol{\mu}) = C(\boldsymbol{\mu})$, demonstrating asymptotic optimality of `KLUCB+Greedy` within these classes. As auxiliary results, we also derive lower bounds on the *sample complexity* of any family of (ε, δ)-algorithms as δ vanishes and show that the sample complexity of `KLUCB+Greedy` closely matches that of the lower bound on the sample complexity.

1 Introduction

Multi-Armed Bandits (MAB) serve as a fundamental framework for sequential decision-making under uncertainty. Traditionally, MAB problems are categorized into two main objectives: minimizing cumulative regret over a finite time horizon, or accurately identifying the best arm.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this work, we study the ε-Best Arm Identification (BAI) problem in the fixed-confidence setting, aiming to output an arm whose mean is within ε of the optimal with probability at least $1-\delta$. Choosing an ε-best arm, rather than the absolute best, is often highly practical, especially when the gap among the top arms is small. In such cases, investing substantial resources (additional pulls and extended testing) to pinpoint the single best arm can be inefficient. An ε-best arm offers a pragmatic balance, delivering near-optimal performance while reducing the costs of exploration.

While prior research on ε-BAI has primarily focused on minimizing *sample complexity*, i.e., the expected number of pulls needed to achieve a desired confidence level, many real-world applications must also contend with the *regret* accrued during the identification process. Our stance is explicitly *regret-aware*: we seek algorithms that return an ε-best arm while minimizing cumulative regret throughout identification. This problem setting aligns with the requirements of many real-world applications, as detailed below.

Clinical Trials. In adaptive clinical trials, each pull assigns a patient to a treatment with initially unknown efficacy. Thus, regret is not just a technical metric: it measures avoidable exposure to inferior treatments while the trial is learning. A regret-aware ε-BAI design aims to recommend a near-optimal therapy while keeping this cumulative harm small.

Motivated by ε-BAI applications where cumulative regret is a primary concern, we revisit ε-BAI from a regret-centric viewpoint. We formulate an ε-BAI problem with minimal regret, in which the learner must, with probability at least $1-\delta$, identify an ε-best arm—i.e., an arm whose mean is at most ε below the optimal mean—while the performance is measured by the expected cumulative regret up to the stopping time. Our main contributions are as follows:

1. In Section 4.2, we establish a fundamental lower bound on the regret for the ε-BAI problem. Specifically, we show that any family of ε-PAC policies indexed by the failure probability $\delta \downarrow 0$ (see its definition in Definition 4.1) must incur an expected cumulative regret up to the stopping time that scales as $\log(1/\delta)$, with an instance-dependent constant $C(\boldsymbol{\mu})$ where $\boldsymbol{\mu}$ is the vector of means. The proof of this lower bound relies on a *regret-aware* change-of-measure technique that allocates information under a fixed regret constraint.
2. In the ε-BAI problem, an algorithm that outputs *any* ε-best arm is acceptable. However, in many applications it is natural to prefer returning an arm with as high a mean reward as possible, since higher reward directly translates into higher profit. We provide several results on this question. Specifically, in Definition 4.4, we define the class of ε-PAC* policies that can return the best arm with probability converging to 1 as $\delta \downarrow 0$, and in Theorem 4.5 we establish a fundamental regret lower bound over this class; see Remark 4.6 for further motivations and justifications for this class of policies. We prove that any ε-PAC* policy must suffer expected cumulative regret up to the stopping time of order $\log(1/\delta)$, with leading constant $C^*(\boldsymbol{\mu})$ that is at least as large as $C(\boldsymbol{\mu})$.
3. In Section 5, we propose `KLUCB+Greedy`. At each round, `KLUCB+Greedy` chooses between the arm with the largest KL-UCB index and the arm with the largest empirical mean. We provide a comprehensive analysis of the correctness and regret of `KLUCB+Greedy`. We show that `KLUCB+Greedy` returns an ε-best arm with probability at least $1-\delta$, achieves expected cumulative regret of order $C^*(\boldsymbol{\mu})\log(1/\delta) + o(\log(1/\delta))$, and belongs to some ε-PAC* family of policies. `KLUCB+Greedy` has two appealing benefits: (i) it is computationally efficient because $C^*(\boldsymbol{\mu})$ admits a closed-form expression; and (ii) `KLUCB+Greedy` is asymptotically optimal for broad classes of reward distributions such as Gaussian distributions, i.e., we have $C^*(\boldsymbol{\mu}) = C(\boldsymbol{\mu})$.

2 Related Work

Best Arm Identification. The fixed-confidence ε-BAI problem, commonly referred to as the *PAC setting*, has been a central focus in multi-armed bandit research. Early foundational work by Even-Dar et al. [11] established initial results for rewards bounded within $[0,1]$, proposing an algorithm with a sample complexity of $O\big((K\log(1/\delta))/\varepsilon^2\big)$. This upper bound was shown to be worst-case optimal by the lower bound derived in Mannor and Tsitsiklis [25]. Subsequent research extended these results to settings where algorithms return k arms that are ε-close to the top-k arms, examining various quality metrics and achieving optimal guarantees—see works such as Kalyanakrishnan and Stone [19], Zhou et al. [32], Cao et al. [4], and Chen et al. [8]. More recently, there has been an emphasis on deriving instance-dependent sample complexity bounds. Algorithms proposed by Karnin et al. [20] and Chen et al. [7] achieve complexities on the order of $O\big(\log(\log(\max\{\Delta_i,\varepsilon\}^{-1})/\delta)/\max\{\Delta_i^2,\varepsilon^2\}\big)$, where

Δ_i denotes the gaps between arm means. Degenne and Koolen [10] analyzed the asymptotic sample complexity, providing matching upper and lower bounds that clarify the fundamental difficulty of the problem. More recently, Jourdan et al. [18] employed the Top-Two algorithm within the ε-BAI framework, achieving an asymptotically optimal solution. Their algorithm additionally supports anytime stopping and offers strong guarantees regarding the quality of the identified arm.

Another line of research focuses on *exact* (i.e., $\varepsilon = 0$) BAI under fixed confidence. Notably, Garivier and Kaufmann [13] provided a comprehensive asymptotic characterization of the sample complexity, achieved through a "tracking" strategy. Recently, Yang et al. [31] investigated minimal regret in exact BAI and established results that are asymptotically optimal.

Finally, we note that BAI can also be studied in the fixed-budget setting, which differs fundamentally from the fixed-confidence framework. The primary metric here is the error probability after a predetermined exploration budget. Carpentier and Locatelli [6] established a lower bound indicating that the algorithms proposed by Audibert et al. [3] and Karnin et al. [20] are already optimal for certain instances. However, a comprehensive asymptotic characterization of the optimal error probability remains largely unresolved and continues to be an active area of research; see Qin [27], Degenne [9], and Wang et al. [30].

Regret Minimization. The study of cumulative regret minimization has long focused on two primary performance metrics: asymptotic regret and worst case regret. Numerous algorithms, such as `KL-UCB` [5] and Thompson Sampling [1, 21, 15–17], have been shown to achieve asymptotic optimality across diverse environments. In terms of worst case regret, the `MOSS` algorithm [2] was the first proven to attain minimax optimality. Subsequently, `KL-UCB`$^{++}$ [26] unified these guarantees by simultaneously attaining both asymptotic and minimax optimality for exponential family distributions.

3 Problem Definition

We consider an agent interacting with K arms. At each time t, the agent selects an arm $A_t \in [K] = \{1, \dots, K\}$ based on past observations and rewards $(A_1, r_1, \dots, A_{t-1}, r_{t-1})$, and receives a reward r_t. The rewards are drawn independently from distributions with unknown means $\mu_1, \dots, \mu_K$. Let $\mathcal{F}_t = \sigma(A_1, r_1, \dots, A_t, r_t)$ denote the σ-field generated by the history up to time t. A policy π is formally defined by a tuple $(\tau, i_{\text{out}}, (\pi_t)_{t=1}^{\infty})$, where:

- $(\pi_t)_{t=1}^{\infty}$ is a sequence of functions and each π_t is $\mathcal{F}_{t-1}$-measurable. For each $t \geq 1$, π_t maps the history up to time $t-1$ to a probability distribution over $[K]$. The arm A_t pulled at time t is sampled according to the distribution $\pi_t(\mathcal{F}_{t-1})$ yielding observation/reward r_t;
- τ is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t=1}^{\infty}$;
- $i_{\text{out}} \in [K]$ is an $\mathcal{F}_\tau$-measurable random variable representing the policy's guess of an ε-best arm.

Definition 3.1 ((ε, δ)-PAC)**.** Fix $\delta \in (0, 1)$. A policy π is (ε, δ)**-probably approximately correct** (or (ε, δ)**-PAC** in short) if for any $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K) \in \mathbb{R}^K$,

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\mu_{i_{\text{out}}} \geq \max_{i \in [K]} \mu_i - \varepsilon\right) \geq 1 - \delta, \qquad \mathbb{P}_{\boldsymbol{\mu}}(\tau < \infty) = 1.$$

We are interested in (ε, δ)-PAC policies that minimize the (expected) regret up to the stopping time τ. We assume $\mu_1 > \max_{i \in [K] \setminus \{1\}} \mu_i$, i.e., arm 1 is the unique optimal arm. Given a policy π that stops at time τ, we define the (cumulative) regret as: $R_{\boldsymbol{\mu}}(\tau) = \sum_{t=1}^{\tau} (\mu_1 - \mu_{A_t})$. The regret represents the cumulative difference between the mean of the optimal arm and the mean of the arm pulled at each step, summed over the duration of the policy.

4 Lower Bounds

In this section, under the fixed-confidence setting with exponential family reward distributions, we present fundamental performance limits for the ε-BAI problem. We begin by formally defining the canonical one-dimensional exponential family and summarizing its key properties. We then establish general lower bounds on regret.[2] Next, we derive regret lower bounds under additional practical constraints on the returned arm. Finally, we present an important impossibility result, revealing an intrinsic trade-off between achieving optimal regret and optimal sample complexity simultaneously.

[2]Throughout this section (or paper), "regret" refers to the expected cumulative regret up to the stopping time, and "sample complexity" refers to the expected number of samples (i.e., the expected stopping time), unless stated otherwise.

4.1 One-Dimensional Exponential Family

We assume each arm follows a canonical one-parameter exponential family $\{\nu_\theta\}_{\theta\in\Theta}$ dominated by ρ on $\mathbb{R}$, with $\frac{d\nu_\theta}{d\rho}(x) = \exp(\theta x - b(\theta))$, where $b(\theta) = \log\int_{\mathbb{R}} e^{\theta x}\,\mathrm{d}\rho(x)$, is the log-partition function and and $\Theta = \{\theta \in \mathbb{R} : b(\theta) < \infty\}$. We make the usual assumption [12, 26, 16] that the log-partition function $b(\theta)$ is twice continuously differentiable. Then, $b'(\theta) = \mathbb{E}_{\nu_\theta}[X] = \mu$ and $b''(\theta) = \mathrm{Var}_{\nu_\theta}(X) = V(\mu) > 0$. The map between μ and θ is one-to-one. Therefore, exponential family of distributions can also be parameterized by the mean value $\mu = b'(\theta)$. We note that $b''(\theta) > 0$ for all θ, which implies $b'(\cdot)$ is invertible and its inverse function b'^{-1} satisfies $\theta = b'^{-1}(\mu)$. Let $\mathcal{D} = \mathrm{int}(\Theta)$ denote the set of values that θ takes on and let $I = \{\mu = b'(\theta) : \theta \in \mathcal{D}\}$ denote the mean-parameter space induced by $\mathcal{D}$. For ease of presentation, we assume $[\mu_i - \varepsilon, \mu_i + \varepsilon] \subseteq I$ for all $i \in [K]$.

The Kullback–Leibler (KL) divergence between distributions in the exponential family with parameters θ, θ' is $\mathrm{KL}(\nu_\theta, \nu_{\theta'}) = b(\theta') - b(\theta) - b'(\theta)(\theta' - \theta)$. We define $\mathrm{kl}(\mu, \mu') = \mathrm{KL}(\nu_\theta, \nu_{\theta'})$. Since the mean uniquely determines the distribution within a one-dimensional exponential family, we let $V(\mu)$ denote the variance of the distribution with mean μ. Moreover, for each arm $i \in [K]$, we assume that the variance of its reward distribution is bounded by a constant V, i.e., $0 < \max_{i\in[K]} V(\mu_i) \le V < +\infty$. For our algorithm, we assume that $V \ge \sup_{\mu\in I} V(\mu)$ is known.

Common distributions within this family include Bernoulli (with $V = 1/4$), Gaussian with known variance (with $V = \sigma^2$). For Gaussian distributions with means $\mu, \mu' \in \mathbb{R}$ and variance $\sigma^2 > 0$, $\mathrm{kl}(\mu, \mu') = (\mu - \mu')^2/(2\sigma^2)$. For Bernoulli distributions with means $\mu, \mu' \in [0, 1]$, $\mathrm{kl}(\mu, \mu') = \mu\log(\mu/\mu') + (1 - \mu)\log((1 - \mu)/(1 - \mu'))$.

We use the canonical one-dimensional exponential-family model following Garivier and Cappé [12], Harremoës [14], Jin et al. [16], Ménard and Garivier [26], Cappé et al. [5]; broader sufficient-statistic or bounded-support models such as Korda et al. [23], Lai and Robbins [24], Riou and Honda [28] are outside our scope.

Frequently Used Notations. We let K be the number of arms. Without loss of generality, we assume that $\mu_1 = \max_{i\in[K]} \mu_i$ throughout this paper. We use Δ_i to denote the gap between arm 1 and arm i, i.e., $\Delta_i := \mu_1 - \mu_i$, $i \in [K] \setminus \{1\}$. We denote $N_i(t) := \sum_{j=1}^{t} \mathbb{1}\{A_j = i\}$ as the number of times that arm i has been played at time step t, and $\widehat{\mu}_i(t) := \sum_{j=1}^{t} \mathbb{1}\{A_j = i\} \cdot r_j/N_i(t)$ as the average reward for pulling arm i up to time t, where r_j is the reward received by the algorithm at time j. Additionally, throughout this paper, we adopt standard asymptotic notations, including little-o, big-O, little omega, and big omega, always with respect to the confidence level δ (tending to zero) , with $\boldsymbol{\mu}$, K, and ε fixed.

4.2 Lower Bound on Regret

Let $\mathcal{I}_\varepsilon(\boldsymbol{\mu}) = \{j \mid \mu_j \ge \max_{i\in[K]} \mu_i - \varepsilon\}$ be the set of arms whose mean is at least $\mu_1 - \varepsilon$. For any candidate arm $j \in [K]$, let $\mathrm{Alt}_j(x) = \big\{\boldsymbol{\lambda} \in I^K : \lambda_1 = x, j \notin \mathcal{I}_\varepsilon(\boldsymbol{\lambda})\big\}$ be the set of bandit instances in which j is not an ε-best arm and the mean of the first arm is x. Define the *regret simplex*

$$\mathcal{W} \;=\; \Big\{\boldsymbol{w} \in \mathbb{R}_+^K :\; w_1 = 0,\; \sum_{i=2}^{K} w_i = 1\Big\}. \tag{1}$$

To study the asymptotic performance of the (ε, δ)-policies as $\delta \downarrow 0$, we consider the *family* of (ε, δ)-PAC policies indexed by $\delta \in (0, 1)$.

Definition 4.1. For $\varepsilon \ge 0$, $\{(\{\pi_{t,\delta}\}_{t=1}^{\infty}, i_{\mathrm{out},\delta}, \tau_\delta)\}_{\delta\in(0,1)}$ is known as a **family of ε-PAC policies** (indexed by $\delta \in (0, 1)$) if each $(\{\pi_{t,\delta}\}_{t=1}^{\infty}, i_{\mathrm{out},\delta}, \tau_\delta)$ is an (ε, δ)-PAC policy.

Theorem 4.2. *For any ε-PAC family of policies,*

$$\liminf_{\delta\downarrow 0} \frac{\mathbb{E}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} \;\ge C(\boldsymbol{\mu}) = \min_{j\in\mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_j(\boldsymbol{\mu}), \tag{2}$$

where

$$\frac{1}{C_j(\boldsymbol{\mu})} = \sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}_j(\mu_1)} \sum_{i=2}^{K} w_i\, \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i}. \tag{3}$$

We defer the proof of Theorem 4.2 to Appendix C. In the proof of Theorem 4.2, we adapt the witness/mixture change-of-measure from Degenne and Koolen [10], but replace sample–allocation and stopping-time reasoning (suited to sample complexity) with a *regret allocation* view—optimizing Kullback–Leibler information per unit regret on a regret simplex. Two pivots make this work: (i) equalization/duality identities that pin down the optimal regret-aware witness under multiple ε-best arms; (ii) a regret-indexed martingale control that bounds log-likelihood ratio fluctuations. This resolves the sample complexity-vs-regret mismatch and yields the lower bound in Theorem 4.2.

On the Characteristic $C_j(\boldsymbol{\mu})$: Theorem 4.2 expresses the regret lower bound through the *characteristic regret times* $C_j(\boldsymbol{\mu})$. To make the bound interpretable and algorithmically useful, we derive a closed-form expression for $j = 1$.

Proposition 4.3 (Closed form for $j=1$)**.** *For a one-parameter exponential family and any instance $\boldsymbol{\mu}$,*

$$C_1(\boldsymbol{\mu}) \;=\; \sum_{i=2}^{K} \frac{\Delta_i}{\mathrm{kl}\big(\mu_i, \mu_1 + \varepsilon\big)}.$$

Since computing $C_j(\boldsymbol{\mu})$ for $j \neq 1$ is involved and not needed later, we defer the details to Appendix E.

4.3 Lower Bound on Regret with Additional Constraint

In the ε-BAI setting, any ε-best arm is an admissible output. However, in many applications it is natural to prefer an arm with the largest mean reward, since higher reward directly translates into higher profit, e.g., in business settings (also see Remark 4.6). Motivated by this, in this paper, we also consider the following family of policies which we call ε-PAC* policies.

Definition 4.4. Fix $\varepsilon > 0$. A family of ε-PAC policies $\{(\{\pi_{t,\delta}\}_{t=1}^{\infty}, i_{\mathrm{out},\delta}, \tau_\delta)\}_{\delta\in(0,1)}$ is known as a **family of ε-PAC* policies** if in addition to the properties of being an ε-PAC (per Definition 4.1),

$$\lim_{\delta\downarrow 0} \mathbb{P}\big(i_{\mathrm{out},\delta} = 1\big) = 1, \tag{4}$$

that is, the probability that the output arm is the best arm converges to 1 as $\delta \downarrow 0$.

Theorem 4.5. *For any ε-PAC* family of policies,*

$$\liminf_{\delta\downarrow 0} \frac{\mathbb{E}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} \geq C^*(\boldsymbol{\mu}) = C_1(\boldsymbol{\mu}). \tag{5}$$

Since $C(\boldsymbol{\mu}) = \min_{j\in\mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_j(\boldsymbol{\mu}) \leq C_1(\boldsymbol{\mu}) = C^*(\boldsymbol{\mu})$, the additional requirement can give a larger lower bound.

Remark 4.6. At first sight, one may wonder whether characterizing the regret complexity for the class of ε-PAC* policies is essentially equivalent to solving the exact BAI problem (with $\varepsilon = 0$), since Definition 4.4 requires the probability of returning the optimal arm to converge to one as $\delta \downarrow 0$. However, we emphasize that ε-PAC* is fundamentally different from exact BAI. Indeed, exact BAI requires the much stronger guarantee $\mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out},\delta} = 1) \geq 1 - \delta$, whereas ε-PAC* only requires the classical statistical notion of *consistency* (of returning the best arm): $\mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out},\delta} = 1) \to 1$ as $\delta \downarrow 0$. Thus, the probability of returning a suboptimal yet still ε-best arm may vanish arbitrarily slowly compared to δ. To make this concrete, consider a family of ε-PAC policies that outputs arm 1 with probability $1 - \sqrt{\delta}$, some other ε-best arm with probability $\sqrt{\delta} - \delta$, and a non-ε-best arm with probability δ. This family is ε-PAC* since $\mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out},\delta} = 1) = 1 - \sqrt{\delta} \to 1$ as $\delta \downarrow 0$, while still satisfying the (ε, δ)-PAC requirement. However, it is not an exact-BAI policy, since the probability of failing to identify the best arm is $\mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out},\delta} \neq 1) = \sqrt{\delta}$, which vanishes strictly more slowly than δ.

More importantly, the class of ε-PAC* policies is well motivated in applications where one wishes to preserve the flexibility of ε-BAI while still preferring higher-quality recommendations whenever possible. For example, in adaptive clinical trials, recommending any treatment whose efficacy is within ε of the best treatment may already be clinically acceptable, especially when treatment differences are small. Nevertheless, as more evidence is collected, one would still prefer the recommendation procedure to increasingly favor the truly best treatment. The ε-PAC* formulation captures precisely this intermediate regime: it allows the learner to exploit the relaxed stopping criterion of ε-BAI, while asymptotically biasing the recommendation toward the truly optimal arm.

In summary, Definition 4.4 can be viewed as a consistency requirement analogous to the classical notion of estimator consistency in asymptotic statistics (cf. van der Vaart [29]): as $\delta \downarrow 0$, the recommendation asymptotically concentrates on the truly optimal arm. At the same time, Definition 4.4 remains fully consistent with the ε-BAI objective. Indeed, even under ε-PAC*, the regret constant becomes $C^*(\boldsymbol{\mu}) = \sum_{i>1} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu_1+\varepsilon)}$, so the slack ε directly lowers the exploration cost relative to exact BAI, while still allowing us to study algorithms that asymptotically favor the optimal arm. $\square$

Since ε-PAC* policies additionally ensure that the probability of returning the best arm converges to 1 as $\delta \downarrow 0$ (a guarantee not required by ε-PAC policies), it is natural to ask how far the complexity $C^*(\boldsymbol{\mu})$, defined as $C_1(\boldsymbol{\mu})$ in (2), can deviate from the constant $C(\boldsymbol{\mu})$, defined in (2).

Lemma 4.7. *Fix an instance $\boldsymbol{\mu}$ and $\varepsilon > 0$. Recall that for a one-parameter exponential family, $V(\mu)$ is the variance of the reward distribution with mean μ. Assume that there exists a constant ℓ such that*

$$\frac{\sup_{\mu \in (\mu_1-\varepsilon, \mu_1+\varepsilon)} V(\mu)}{\inf_{\mu \in (\mu_1-\varepsilon, \mu_1+\varepsilon)} V(\mu)} \leq \ell. \tag{6}$$

Then, $C(\boldsymbol{\mu}) \leq C^*(\boldsymbol{\mu}) \leq \ell C(\boldsymbol{\mu})$.

In fact, for Gaussian rewards, Lemma 4.7 can be strengthened and no gap exists between $C^*(\boldsymbol{\mu})$ and $C(\boldsymbol{\mu})$. This is shown in Lemma E.4 in which we prove that for Gaussian rewards, $C^*(\boldsymbol{\mu}) = C(\boldsymbol{\mu})$. In addition, in Appendix E.1 we quantify the gap between $C^*(\boldsymbol{\mu})$ and $C(\boldsymbol{\mu})$ for Bernoulli rewards on some special instances.

4.4 On the Sample Complexity

It is easy to see that there exists an ε-BAI algorithm that is *asymptotically optimal in sample complexity* (e.g., Degenne and Koolen [10]) and incurs cumulative regret of order $\Theta(\log(1/\delta))$. This naturally raises a question: *does there exist an algorithm that is asymptotically optimal for regret while also achieving $\Theta(\log(1/\delta))$ sample complexity?* In this section we consider this problem for families of ε-PAC* policies.

Theorem 4.8 (Impossibility result)**.** *Fix an instance $\boldsymbol{\mu}$. For any ε-PAC* family of policies, if there exists a constant $\kappa > 0$ such that, for all sufficiently small δ,*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \leq \kappa \log(1/\delta), \quad \textit{then} \quad \liminf_{\delta \downarrow 0} \frac{\mathbb{E}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} > C^*(\boldsymbol{\mu}).$$

Discussion. Theorem 4.8 shows that for the family of ε-PAC* policies, it is not possible to achieve the optimal regret constant $C^*(\boldsymbol{\mu})$ (cf. Theorem 4.5) if the expected number of arm pulls $\mathbb{E}[\tau_\delta]$ is $O(\log(1/\delta))$. Hence, to achieve $C^*(\boldsymbol{\mu})$, $\mathbb{E}[\tau_\delta]$ must be asymptotically larger than $\log(1/\delta)$.

The extra samples come from arm 1: attaining $C^*(\boldsymbol{\mu})$ requires $\mathbb{E}_{\boldsymbol{\mu}}[N_1(\tau_\delta)] = \omega(\log(1/\delta))$, while each suboptimal arm needs only $(1 + o(1)) \log(1/\delta)/\mathrm{kl}(\mu_i, \mu_1 + \varepsilon)$ samples. These best-arm pulls incur no regret and shrink its confidence interval. This with further motivates the greedy component of our algorithm.

5 The Proposed Algorithm: `KLUCB+Greedy`

In this section, we first present a high-level overview of the elements in the design of our algorithm. This then allows us to introduce `KLUCB+Greedy`, after which we provide guarantees on its regret and sample complexity.

5.1 High Level Overview

For ease of explanation, we assume the underlying reward distribution is Gaussian with unit variance, although the analysis extends to other distributions within the one-dimensional exponential family.

At time t, the algorithm considers a candidate for the ε-best arm, denoted by $\mathrm{a}^*(t)$. The algorithm outputs this arm when a suitably defined lower confidence bound (LCB) $L_{\mathrm{a}^*(t)}(t)$ meets the following condition relative to suitably defined upper confidence bounds (UCBs) of all other arms $\{U_i(t)\}_{i \in [K] \setminus \{\mathrm{a}^*(t)\}}$:

$$L_{\mathrm{a}^*(t)}(t) \geq \max_{i \in [K] \setminus \{\mathrm{a}^*(t)\}} U_i(t) - \varepsilon + 2\zeta,$$

for some $\zeta = \zeta_{\varepsilon,\delta}$ to be determined later. The design of these UCBs and LCBs is such that the following guarantees simultaneously hold up to the stopping time with probability at least $1 - \delta/2$:

$$L_{\mathrm{a}^*(t)}(t) \le \mu_{\mathrm{a}^*(t)} + \zeta \quad \text{and} \tag{7}$$

$$U_1(t) \ge \mu_1 - \zeta, \tag{8}$$

where μ_i is the true mean reward of arm i. (7) is guaranteed by Lemma D.2, while (8) follows from Lemma D.1. From (7) and (8), if $\mathrm{a}^*(t) \neq 1$, we have

$$\mu_{\mathrm{a}^*(t)} + \zeta \ge L_{\mathrm{a}^*(t)}(t) \ge \max_{i \in [K] \setminus \{\mathrm{a}^*(t)\}} U_i(t) - \varepsilon + 2\zeta \ge \mu_1 - \varepsilon + \zeta,$$

which indicates that $\mu_{\mathrm{a}^*(t)} \ge \mu_1 - \varepsilon$ and thus $\mathrm{a}^*(t)$ is the ε-best arm. If $\mathrm{a}^*(t) = 1$, the output is already correct.

Recall that $N_i(t)$ is the number of pulls of arm i by time t. For ε-PAC* policies and each suboptimal arm, the lower-bound constraints require $\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau_\delta)] \ge (1 - o(1)) \log(1/\delta)/\mathrm{kl}(\mu_i, \mu_1 + \varepsilon)$. Since regret is the sum of these counts weighted by the positive gaps Δ_i, attaining $C^*(\boldsymbol{\mu}) \log(1/\delta)$ requires each bound to be tight to first order. Assume that $N_1(t) = \omega(\log(1/\delta))$. To design an algorithm with expected regret that scales as $(1 + o(1))C^*(\boldsymbol{\mu}) \log(1/\delta)$, we aim for $U_1(t)$ to satisfy

$$U_1(t) \in [\mu_1 - m(\delta),\, \mu_1 + m(\delta)] \qquad \text{with} \quad m(\delta) \to 0 \text{ as } \delta \downarrow 0. \tag{9}$$

Intuitively, it suffices to certify that every suboptimal arm $i \neq 1$ has mean at most $\mu_1 + \varepsilon$, which implies that arm 1 is ε-best. Under (9), the index $U_1(t)$ concentrates around μ_1. We similarly require $L_1(t) = \mu_1 + o(1)$. A nonvanishing error in the best-arm bounds would reduce the gap used in the stopping rule and increase the leading number of suboptimal-arm pulls. Consequently, in this Gaussian unit-variance heuristic, each suboptimal arm $i \neq 1$ requires on the order of $2\log(K/\delta)/(\Delta_i + \varepsilon)^2$ pulls to certify that $\mu_i < \mu_1 + \varepsilon$ and the algorithm can then return arm 1 as an ε-best arm.

We claim that the properties in (8) and (9) are realized by sampling arm 1 sufficiently often and defining the UCB and LCB respectively as $U_i(t) = \widehat{\mu}_i(t) + \sqrt{\frac{2f_{N_i(t)}(1)}{N_i(t)}}$, and $L_i(t) = \widehat{\mu}_i(t) - \sqrt{\frac{2f_{N_i(t)}(K)}{N_i(t)}}$, where $f_s(x)$ is defined as

$$f_s(x) = \log\Big(\frac{Qx\log^2(1/\bar{\delta})}{\min\{1, s\varepsilon^2/V\}\bar{\delta}}\Big), \tag{10}$$

with $Q = 128e^4$. The parameters $\bar{\delta}$ and ζ are specified in (11); $\bar{\delta} = \delta$ for all sufficiently small δ. This can be seen by applying appropriate concentration inequalities to show that these functional forms satisfy (8) and (9). The factor K in $f_{N_i(t)}(K)$ for $L_i(t)$ is included to accommodate a union bound over all K arms, ensuring $L_{\mathrm{a}^*(t)}(t) \le \mu_{\mathrm{a}^*(t)} + \zeta$ holds with high probability for any candidate arm.

For Gaussian rewards with known variance σ^2, the algorithm uses $\mathrm{kl}(x, y) = (x - y)^2/(2\sigma^2)$. The two confidence radii above are multiplied by σ, and $V = \sigma^2$. The sampling rule is unchanged.

Greedy pulls. To obtain these extra samples from the best arm, we introduce a novel idea. Specifically, we introduce a *candidate set* of arms $\mathcal{C}_t$ defined as

$$\mathcal{C}_t = \left\{ i \in [K] : \widehat{\mu}_i(t) \ge \max_{j \in [K]} U_j(t) - \varepsilon + \frac{7\zeta}{2} \right\}$$

where ζ is specified in (11). We then select $A_{t+1} = \mathrm{a}^*(t)$ if $\mathcal{C}_t \neq \emptyset$, and choose $A_{t+1} = \arg\max_{i \in [K]} U_i(t)$ otherwise. When $\mathcal{C}_t \neq \emptyset$, we update the candidate to the arm with the largest empirical mean in $\mathcal{C}_t$ only if its empirical mean exceeds that of the previous candidate by at least ζ. When $\max_i U_i(t) = \mu_1 + \varepsilon + O(\zeta)$, the threshold defining $\mathcal{C}_t$ is $\mu_1 + O(\zeta)$. Thus the candidate set consists of arms whose empirical means are close to the optimal mean.

Why does this work? When $\max_i U_i(t) \le \mu_1 + \varepsilon - 5\zeta$ and $\widehat{\mu}_1(t) \ge \mu_1 - \zeta$, the optimal arm belongs to $\mathcal{C}_t$: the threshold defining $\mathcal{C}_t$ is at most $\mu_1 - 3\zeta/2$. The algorithm then starts favoring the current candidate. Upon stopping, although the optimal arm has been pulled $\omega(\log(1/\delta))$ times and $U_1(t) \approx \mu_1$, the upper confidence bounds of suboptimal arms remain around $\mu_1 + \varepsilon - \Theta(\zeta)$, ensuring $(1 + o(1))C^*(\boldsymbol{\mu}) \log(1/\delta)$ regret.

At finite δ, a greedy pull can still select a suboptimal arm. Our analysis shows that these pulls contribute $o(\log(1/\delta))$ expected regret. The role of ζ is to leave a small margin in the confidence and switching rules, while tending to zero as $\delta \downarrow 0$.

KL-UCB fails to achieve $(1+o(1))C^*(\boldsymbol{\mu})\log(1/\delta)$ regret. One may wonder whether the classic KL-UCB which is asymptotically optimal for classical regret minimization [12] is also asymptotically optimal for our ε-BAI setting, if we use an appropriate stopping rule. Recall that KL-UCB implements the naïve rule that pulls the arm with the largest index at each round, $A_{t+1} \in \arg\max_{i\in[K]} U_i(t)$, where $U_i(t) = \sup\{\widetilde{\mu} \in I : \mathrm{kl}(\widehat{\mu}_i(t), \widetilde{\mu}) \leq \frac{h(t)}{N_i(t)}\}$ and $h(t) = \log(1+t\log^2 t)$. Let $t_0 = \log^2(1/\delta)$. Consider the time $t_1 > t_0$. Since at time t_1, the regret of KL-UCB is $\Theta(\log t_1)$, it pulls the best arm on most rounds; in particular $N_1(t_1) \gg h(t_1)$ and $\widehat{\mu}_1(t_1) = \mu_1 + o(1)$ (we omit "with high probability" for simplicity). Consequently, $U_1(t_1) = \mu_1 + o(1)$. Under the selection rule $A_{t+1} \in \arg\max_i U_i(t)$, by time t_1 each suboptimal arm $i \neq 1$ must have its UCB to be close to arm 1's UCB; otherwise KL-UCB would keep selecting arm i and increasing $N_i(t)$ until $U_i(t)$ decreases to be near to $U_1(t)$. Hence all $\{U_i(t_1)\}_{i\in[K]\setminus\{1\}}$ approach $U_1(t_1)$, and together with the fact that $U_1(t_1) = \mu_1 + o(1)$ this forces, for every $i \neq 1$, $U_i(t_1) = \mu_1 + o(1)$. By the definition of $U_i(t_1)$, this implies that $N_i(t_1)$ must be large enough so that its UCB reaches $\mu_1 + o(1)$, yielding

$$N_i(t_1) \approx \frac{\log(1+t_1\log^2(t_1))}{\mathrm{kl}(\mu_i + o(1), \mu_1 + o(1))}, \qquad \forall\, i \neq 1,$$

where $\mu_i + o(1)$ follows since arm i has been pulled $\Theta(\log t_1)$ times, so its empirical mean concentrates around μ_i. In contrast, achieving regret $(1+o(1))C^*(\boldsymbol{\mu})\log(1/\delta)$ requires

$$N_i[\tau_\delta] \;\approx\; \frac{\log(1/\delta)}{\mathrm{kl}(\mu_i, \mu_1+\varepsilon)}, \qquad \forall\, i \neq 1.$$

If KL-UCB achieves $(1+o(1))C^*(\boldsymbol{\mu})\log(1/\delta)$ regret, KL-UCB would need a stopping time $\tau_\delta = t_1$ such that, simultaneously for all $i \neq 1$,

$$\frac{\log(1+t_1\log^2(t_1))}{\mathrm{kl}(\mu_i, \mu_1)} \;\approx\; \frac{\log(1/\delta)}{\mathrm{kl}(\mu_i, \mu_1+\varepsilon)}.$$

Equivalently, $\log t_1$ would have to satisfy

$$\log t_1 \approx \log(1/\delta) \cdot \frac{\mathrm{kl}(\mu_i, \mu_1)}{\mathrm{kl}(\mu_i, \mu_1+\varepsilon)} \qquad \text{for all } i \neq 1,$$

which is impossible in general since the ratio $\mathrm{kl}(\mu_i, \mu_1)/\mathrm{kl}(\mu_i, \mu_1+\varepsilon)$ depends on i. Hence, KL-UCB fails to achieve $(1+o(1))C^*(\boldsymbol{\mu})\log(1/\delta)$ regret in general.

5.2 The Proposed Algorithm

For $\varepsilon > 0$ and $\delta \in (0,1)$, let

$$L_0 = \max\{32, 2\log K, \varepsilon^4\}, \qquad \bar{\delta} = \min\{\delta, e^{-L_0}\},$$
$$\zeta = \min\left\{\frac{1}{\sqrt{\log\log(1/\bar{\delta})}}, \frac{\varepsilon}{4}\right\}. \tag{11}$$

We have $\bar{\delta} = \delta$ for all sufficiently small δ, so this does not change the asymptotic bounds.
Algorithm 1 presents the pseudocode of the proposed KLUCB+Greedy algorithm for sequential arm selection in a K-armed bandit problem with the objective of being an ε-PAC* family of policies as $\delta \downarrow 0$. KLUCB+Greedy requires ε, δ and V as inputs, and computes ζ and T. The algorithm is initialized by playing each arm once. At each subsequent step t, it computes UCBs, $\{U_i(t)\}_{i\in[K]}$ and identifies a set of candidate arms $\mathcal{C}_t$ based on UCBs. The arm A_{t+1} to be played is selected based on the following rules, which also determines the candidate of the best arm $\mathrm{a}^*(t)$:

1. If the candidate set $\mathcal{C}_t$ is empty, the algorithm plays $A_{t+1} = b_t$ (Line 12) and retains $\mathrm{a}^*(t) = \mathrm{a}^*(t-1)$, where $b_t = \arg\max_{i\in[K]} U_i(t)$ is the arm with the highest UCB.
2. Else, if a greedy condition on empirical means is met, comparing $\widehat{\mu}_{\mathrm{a}^*(t-1)}(t)$ with values related to arms in $\mathcal{C}_t$, the algorithm updates $\mathrm{a}^*(t)$ to be the arm in $\mathcal{C}_t$ with the highest empirical mean ($\arg\max_{i\in\mathcal{C}_t} \widehat{\mu}_i(t)$) and plays this arm, $A_{t+1} = \mathrm{a}^*(t)$.

Algorithm 1 KLUCB+Greedy

1: **Input:** a set of arms $[K]$, ε, δ, known variance bound V
2: **Output:** i_{out}
3: Compute $\bar{\delta}$ and ζ from (11).
4: Set $T = 3K + \lceil 32KV\zeta^{-2}\log\big(QK\log^2(1/\bar{\delta})/\bar{\delta}\big)\rceil$.
5: **Initialization:** Choose each arm once. Let $\mathrm{a}^*(K-1) = \arg\max_i \widehat{\mu}_i(K)$.
6: **for** $t \leftarrow K$ **to** $T-1$ **do**
7: **For each** $i \in [K]$, compute the UCBs
$$U_i(t) = \sup\left\{\mu \in I : \mathrm{kl}(\widehat{\mu}_i(t), \mu) \le \frac{f_{N_i(t)}(1)}{N_i(t)}\right\}$$
8: Let $b_t = \arg\max_{i\in[K]} U_i(t)$.
9: Let $\mathcal{C}_t = \{i : \widehat{\mu}_i(t) \ge U_{b_t}(t) - \varepsilon + 7\zeta/2\}$.
10: **if** $\mathcal{C}_t = \emptyset$ **then**
11: $\mathrm{a}^*(t) = \mathrm{a}^*(t-1)$.
12: Select $A_{t+1} = b_t$. */* Choose the arm with the largest UCB. */*
13: **else if** $\max_{i\in\mathcal{C}_t} \widehat{\mu}_i(t) \ge \widehat{\mu}_{\mathrm{a}^*(t-1)}(t) + \zeta$ **then**
14: $\mathrm{a}^*(t) = \arg\max_{i\in\mathcal{C}_t} \widehat{\mu}_i(t)$.
15: Select $A_{t+1} = \mathrm{a}^*(t)$. */* Choose the arm greedily. */*
16: **else**
17: $\mathrm{a}^*(t) = \mathrm{a}^*(t-1)$.
18: Select $A_{t+1} = \mathrm{a}^*(t)$.
19: **end if**
20: Compute the LCB
$$L_{\mathrm{a}^*(t)}(t) = \inf\left\{\mu \in I : \mu \le \widehat{\mu}_{\mathrm{a}^*(t)}(t),\ \mathrm{kl}(\widehat{\mu}_{\mathrm{a}^*(t)}(t), \mu) \le \frac{f_{N_{\mathrm{a}^*(t)}(t)}(K)}{N_{\mathrm{a}^*(t)}(t)}\right\}.$$
21: **if** $L_{\mathrm{a}^*(t)}(t) \ge \max_{i\neq \mathrm{a}^*(t)} U_i(t) - \varepsilon + 2\zeta$ **then**
22: Return $i_{\text{out}} = \mathrm{a}^*(t)$ with $\tau_\delta = t$.
23: **end if**
24: Play A_{t+1}, observe its reward, and update the sample statistics.
25: **end for**
26: Return $i_{\text{out}} = \arg\max_i \widehat{\mu}_i(T)$ with $\tau_\delta = T$.

3. Else, the algorithm plays $A_{t+1} = \mathrm{a}^*(t-1)$ (Line 18) and retains it as the current candidate of the best arm, $\mathrm{a}^*(t) = \mathrm{a}^*(t-1)$.

After updating the candidate, using the first t observations, an LCB $L_{\mathrm{a}^*(t)}(t)$ is computed for $\mathrm{a}^*(t)$. The algorithm employs a confidence-based stopping rule: it terminates at time $\tau = t$ and returns $\mathrm{a}^*(t)$ if the condition in Line 21 is met, requiring $L_{\mathrm{a}^*(t)}(\tau)$ to be sufficiently separated from the UCBs of all other arms. If it does not stop at time t, it plays A_{t+1} and updates the sample statistics. If this condition is not met before T, the algorithm returns the arm with the highest empirical mean.

Comparison with DKL-UCB **[31].** The DKL-UCB algorithm of Yang et al. [31] is designed for exact BAI. Although it also modifies KL-UCB-style sampling and uses two KL-UCB indices, U^f and U^g, the additional index U^g is introduced to accelerate the learning of suboptimal arms; using only U^f would be too slow and may lead to sample complexity of order $\Theta(1/\delta)$ [31]. In contrast, KLUCB+Greedy uses modified KL-UCB indices that explicitly depend on δ and ε, and its greedy component is used to allocate more pulls to the best arm.

5.3 Main Results

The following theorem, which is proved in Appendix D, demonstrates the correctness of KLUCB+Greedy and establishes upper bounds on its sample complexity and regret.

Theorem 5.1. *Let ζ and $\bar{\delta}$ be as defined in* (11) *and $T = 3K + \lceil 32KV\zeta^{-2}\log\big(QK\log^2(1/\bar{\delta})/\bar{\delta}\big)\rceil$, where $Q = 128e^4$. For any exponential family of distributions $\mathcal{F}$ with maximum variance V, KLUCB+Greedy indexed by the failure probability $\delta \in (0,1)$ is ε-PAC*. Furthermore, its regret*

asymptotically satisfies

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} = C^*(\boldsymbol{\mu}) = \sum_{i=2}^{K} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu_1 + \varepsilon)}, \tag{12}$$

and its sample complexity satisfies

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)(\log\log(1/\delta))^2} = 0. \tag{13}$$

In the following, we discuss the optimality of regret and sample complexity of `KLUCB+Greedy`.

1. Combining Theorem 5.1 with Theorem 4.2 and the discussion following Lemma 4.7, we conclude that among all ε-PAC policies (cf. Definition 4.1), `KLUCB+Greedy` achieves the asymptotic optimal regret for Gaussian rewards with fixed/known variance, and is near asymptotically optimal (in the sense that the upper and lower bounds differ by a constant ℓ in (6)) when the reward variance is approximately constant over the interval $(\mu_1 - \varepsilon, \mu_1 + \varepsilon)$.
2. Combining Theorem 5.1 and Theorem 4.5, we conclude that `KLUCB+Greedy` achieves the optimal regret among all ε-PAC* (cf. Definition 4.4) policies.
3. Although our primary objective is the expected cumulative regret rather than the expected sample complexity, (13) implies that the sample complexity of `KLUCB+Greedy` is also close to the fundamental limit in Theorem 4.8, up to a small multiplicative factor of order $(\log\log(1/\delta))^2$. Consequently, `KLUCB+Greedy` has near-optimal sample complexity.

Yang et al. [31] study exact BAI, corresponding to $\varepsilon = 0$, whereas our ε-PAC* formulation improves the divergence term from $\mathrm{kl}(\mu_i, \mu_1)$ to $\mathrm{kl}(\mu_i, \mu_1 + \varepsilon)$. Moreover, as $\varepsilon \to 0$, $C^*(\boldsymbol{\mu})$ tends to the regret constant of the `DKL-UCB` algorithm in Yang et al. [31].

Experiments. Due to space limitations, all experiments, which strongly corroborate the theory, are deferred to Appendix A.

6 Conclusion and Future Work

This work studied fixed-confidence ε-BAI with a focus on *minimizing regret*. We proved the fundamental lower bounds for ε-BAI: any family of ε-PAC policies (cf. Definition 4.1) suffers a regret of at least $C(\boldsymbol{\mu})\log(1/\delta)(1 + o(1))$ and any family of ε-PAC* policies (cf. Definition 4.4) suffers a regret of at least $C^*(\boldsymbol{\mu})\log(1/\delta)(1 + o(1))$. We also showed a trade-off—to achieve the optimal regret among all families of ε-PAC* policies, the sample complexity must be $\omega(\log(1/\delta))$.

To achieve the lower bound for ε-PAC* policies, we proposed `KLUCB+Greedy`, which we prove belongs to the class of ε-PAC* policies. It achieves regret $(C^*(\boldsymbol{\mu}) + o(1))\log(1/\delta)$ and sample complexity $o\big(\log(1/\delta)(\log\log(1/\delta))^2\big)$. `KLUCB+Greedy` is optimal in regret among ε-PAC* policies and nearly optimal in sample complexity within that class.

Several questions remain open. In particular, it is interesting to ask whether one can design an algorithm with asymptotically optimal regret among all ε-PAC policies. The main challenge is to learn the least costly valid output and its sampling allocation. We surmise that adapting track-and-stop–type techniques [13, 10] to our setting could achieve this goal. A finite-confidence regret bound with explicit lower-order terms is also open.

Acknowledgment

Tianyuan Jin is partially supported by National Key R&D Program of China under Grant No. 2024YFA1012700. V. Y. F. Tan is supported by a Singapore Ministry of Education AcRF Tier 2 grant (A-8004062-00-00) and a Singapore National Research Foundation Investigatorship (A-8004657-00-00).

References

[1] Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017. 3

[2] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009. 3

[3] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 41–53, 2010. 3

[4] Wei Cao, Jian Li, Yufei Tao, and Zhize Li. On top-k selection in multi-armed bandits and hidden bipartite graphs. In *Advances in Neural Information Processing Systems*, pages 1036–1044, 2015. 2

[5] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013. 3, 4, 13

[6] 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. 3

[7] Jiecao Chen, Xi Chen, Qin Zhang, and Yuan Zhou. Adaptive multiple-arm identification. In *International Conference on Machine Learning*, pages 722–730, 2017. 2

[8] Lijie Chen, Anupam Gupta, and Jian Li. Pure exploration of multi-armed bandit under matroid constraints. In *Conference on Learning Theory*, pages 647–669, 2016. 2

[9] 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. 3

[10] Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 3, 5, 6, 10, 22

[11] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002. 2

[12] 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, 2011. 4, 8

[13] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016. 3, 10, 13, 38

[14] Peter Harremoës. Bounds on tail probabilities in exponential families. *arXiv preprint arXiv:1601.05179*, 2016. 4, 15

[15] Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021. 3

[16] Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Anima Anandkumar. Finite-time regret of Thompson sampling algorithms for exponential family multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 35, pages 38475–38487, 2022. 4

[17] Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration is fast and optimal. In *International Conference on Machine Learning*, pages 15239–15261. PMLR, 2023. 3

[18] Marc Jourdan, Rémy Degenne, and Emilie Kaufmann. An ε-best-arm identification algorithm for fixed-confidence and beyond. In *Advances in Neural Information Processing Systems*, volume 36, pages 16578–16649, 2023. 3

[19] Shivaram Kalyanakrishnan and Peter Stone. Efficient selection of multiple bandit arms: Theory and practice. In *ICML*, pages 511–518, 2010. 2

[20] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246, 2013. 2, 3

[21] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012. 3

[22] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1): 1–42, 2016. 21

[23] Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013. 4

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

[25] Shie Mannor and John N. Tsitsiklis. Lower bounds on the sample complexity of exploration in the multi-armed bandit problem. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24–27, 2003. Proceedings*, pages 418–432. Springer, 2003. 2

[26] Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237, 2017. 3, 4, 15

[27] Chao Qin. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, pages 5650–5654. PMLR, 2022. 3

[28] Charles Riou and Junya Honda. Bandit algorithms based on Thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pages 777–826. PMLR, 2020. 4

[29] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. 6

[30] Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Best arm identification with fixed budget: A large deviation perspective. In *Advances in Neural Information Processing Systems*, volume 36, pages 16804–16815, 2023. 3

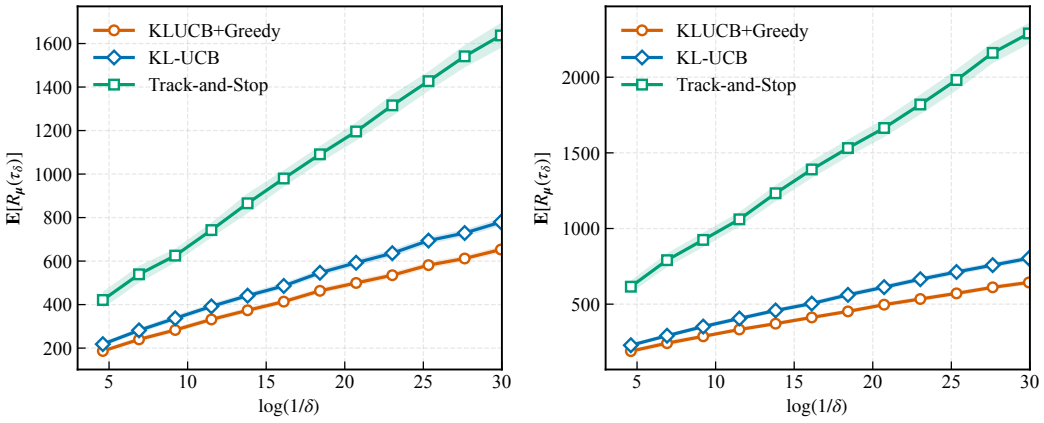
[31] Junwen Yang, Vincent Y. F. Tan, and Tianyuan Jin. Best arm identification with minimal regret. *Journal of Machine Learning Research*, 27(86):1–41, 2026. URL `https://jmlr.org/papers/v27/24-1612.html`. 3, 9, 10, 20

[32] Yuan Zhou, Xi Chen, and Jian Li. Optimal PAC multiple arm identification with applications to crowdsourcing. In *International Conference on Machine Learning*, pages 217–225, 2014. 2

Supplementary Material for the NeurIPS 2026 Submission: On Minimizing Regret in Fixed-Confidence ε-Best Arm Identification

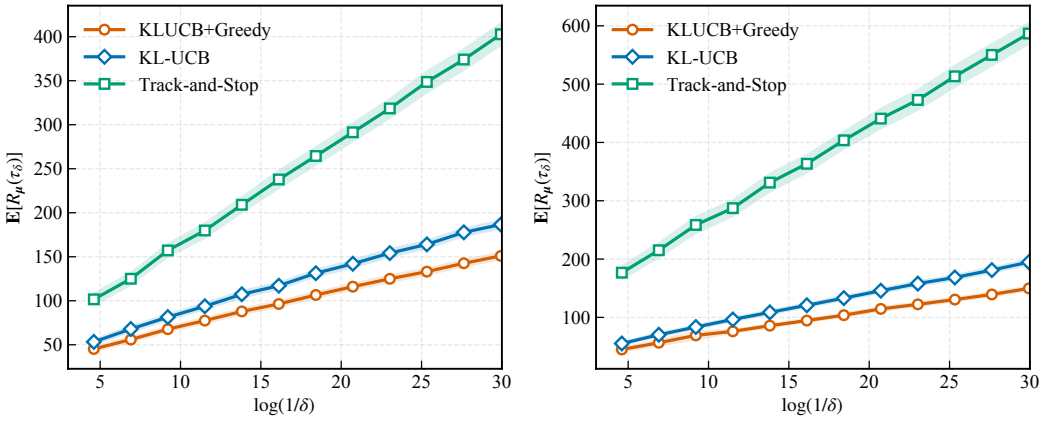
A Experiments

We evaluate the *expected cumulative regret up to stopping time*. Each experiment is averaged over 500 independent trials. The experiments use the original confidence parameters; the finite-confidence guarantees in Theorem 5.1 apply to the revised parameters. We compare `KLUCB+Greedy` with `KL-UCB` [5] and `Track-and-Stop` [13]. All our experiments are implemented in MATLAB and parallelized using an Intel® Xeon® Gold 6244 CPU (3.60 GHz) with 32 GB memory.



(a) $\boldsymbol{\mu}_1 = (0.6, 0.3, 0.3, 0.3, 0.3, 0.3)$ (b) $\boldsymbol{\mu}_2 = (0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$

Figure 1: The regret for different algorithms on Gaussian rewards. The experiments are averaged over 500 repetitions and the error bars denote one standard deviation above/below the mean result.



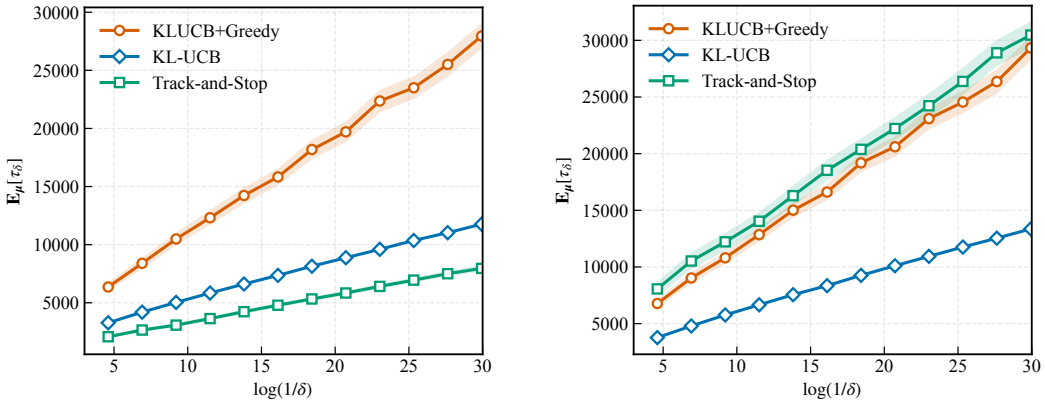
(a) $\boldsymbol{\mu}_1 = (0.6, 0.3, 0.3, 0.3, 0.3, 0.3)$ (b) $\boldsymbol{\mu}_2 = (0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$

Figure 2: The regret for different algorithms on Bernoulli rewards. The experiments are averaged over 500 repetitions.

Setup. We set the suboptimality parameter $\varepsilon = 0.25$ and consider two 6-arm bandit instances (i.e., $K = 6$) with mean vectors $\boldsymbol{\mu}_1 = (0.6, 0.3, 0.3, 0.3, 0.3, 0.3)$ and $\boldsymbol{\mu}_2 = (0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$. In both instances, arm 1 is optimal, and any ε-best arm satisfies $\mu_i \geq \mu_1 - \varepsilon = 0.6 - 0.25 = 0.35$. Hence, for $\boldsymbol{\mu}_1$ there is a unique ε-best arm, whereas for $\boldsymbol{\mu}_2$ there are multiple ε-best arms, namely those with means 0.6, 0.5, and 0.4. Throughout the experiments, rewards are drawn independently either from a Gaussian distribution with unit variance or from a Bernoulli distribution. In each case, a bandit instance is fully specified by its mean vector $\boldsymbol{\mu}$, which is reported below the figures. For `KL-UCB`, the algorithm is not designed for ε-BAI and does not have stopping rule. Therefore, when running `KL-UCB` in our experiments, we couple it with the same stopping rule as `KLUCB+Greedy`: after pulling an arm at round t, we compute the UCBs (of the same form as those in Line 7 of Algorithm 1) and the LCBs (of the same form as those in Line 20 of Algorithm 1) for all arms, and then check whether the stopping condition is satisfied for some arm i, namely, $L_i(t) \geq \max_{j \in [K] \setminus \{i\}} U_j(t) - \varepsilon + 2\zeta$.

Results on Regret. Across all our experiments, all compared algorithms consistently identify the true best arm, i.e., they always return the optimal arm at termination even when there are multiple arms as in instance $\boldsymbol{\mu}_2$. This aligns with the fact that `KLUCB+Greedy` is ε-PAC*.

Moreover, as shown in Figures 1 and 2, `KLUCB+Greedy` achieves the smallest expected regrets among the compared methods, which demonstrates the effectiveness of our proposed algorithm. Furthermore, `KL-UCB` exhibits larger regret than `KLUCB+Greedy`, consistent with the discussion in Section 5.1, where we show that `KL-UCB` cannot attain the same leading regret constant as `KLUCB+Greedy`, namely $C^*(\boldsymbol{\mu}) \log(1/\delta) + o(\log(1/\delta))$. Finally, `Track-and-Stop` incurs substantially larger regret than both `KLUCB+Greedy` and `KL-UCB` under Gaussian and Bernoulli rewards, which is expected since it is designed to minimize sample complexity rather than cumulative regret.



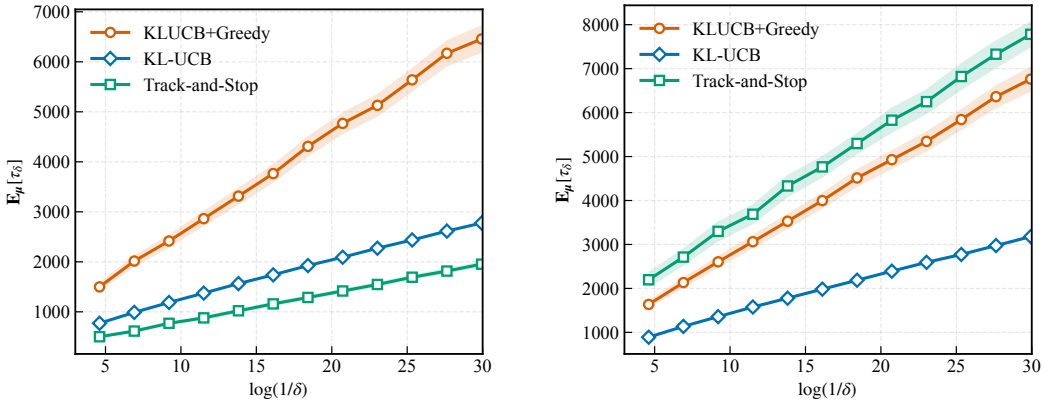
(a) $\boldsymbol{\mu} = (0.6, 0.3, 0.3, 0.3, 0.3, 0.3)$ (b) $\boldsymbol{\mu} = (0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$

Figure 3: The sample complexity for different algorithms on Gaussian rewards. The experiments are averaged over 500 repetitions.

Results on Sample Complexity. The results are shown in Figures 3 and 4.

For the easier instance $\boldsymbol{\mu} = (0.6, 0.3, 0.3, 0.3, 0.3, 0.3)$, `Track-and-Stop` achieves the smallest sample complexity, while `KL-UCB` requires more samples and `KLUCB+Greedy` has the largest sample complexity. This is because, unlike `Track-and-Stop`, which aims to minimize the sample complexity, `KLUCB+Greedy` and `KL-UCB` aim to minimize the regret.

For the instance $\boldsymbol{\mu} = (0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$, `KL-UCB` has the smallest sample complexity, whereas `Track-and-Stop` has the largest sample complexity and `KLUCB+Greedy` lies in between. This is because `Track-and-Stop` must return the best arm with probability $1 - \delta$ and the gap of the best arm and second best arm is $0.1 < \varepsilon = 0.25$, whereas `KLUCB+Greedy` and `KL-UCB` terminate once the stopping criterion guaranteeing an ε-best arm is satisfied. This demonstrates that although `KLUCB+Greedy` is designed to belong to ε-PAC* policies (whose success probability of returning the best arm tends to 1 as $\delta \downarrow 0$), it can still benefit from allowing the output to be any ε-best arm rather than the unique best arm.



(a) $\boldsymbol{\mu} = (0.6, 0.3, 0.3, 0.3, 0.3, 0.3)$ (b) $\boldsymbol{\mu} = (0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$

Figure 4: The sample complexity for different algorithms on Bernoulli rewards. The experiments are averaged over 500 repetitions.

B Useful Inequalities and Equalities

The following key properties of the KL divergence are frequently used in the proof:

Proposition B.1 (Harremoës [14]). $\mathrm{kl}(\mu, \mu') = \int_{\mu}^{\mu'} \frac{x-\mu}{V(x)}\, dx$.

Proposition B.2. *For* $\Delta > 0$ *and* $\mu \leq \mu' - \Delta$, $\mathrm{kl}(\mu, \mu') \geq \mathrm{kl}(\mu, \mu' - \Delta)$ *and* $\mathrm{kl}(\mu, \mu') \leq \mathrm{kl}(\mu - \Delta, \mu')$.

Lemma B.3 (Pinsker's Type Inequality). *Assume* P *and* Q *are distributions that follow some natural exponential family. Let* μ_P *and* μ_Q *be the means of* P *and* Q. *Assume that* $V \geq \sup_{x \in [\mu_P \wedge \mu_Q,\, \mu_P \vee \mu_Q]} V(x)$. *Then,*

$$\mathrm{KL}(P, Q) \geq \frac{(\mu_P - \mu_Q)^2}{2V}.$$

We also introduce the following concentration bounds.

Lemma B.4 (Maximal Inequality; Lemma 4 of Ménard and Garivier [26]). *Let* N *and* M *be two positive integers, and let* $\widehat{\mu}_n$ *be the empirical mean of* n *i.i.d. random variables from a natural exponential-family distribution with mean* μ *and variance upper bound* V. *Then, for* $x \leq \mu$,

$$\begin{aligned} \mathbb{P}(\exists N \leq n \leq M, \widehat{\mu}_n \leq x) &\leq e^{-N \cdot \mathrm{kl}(x,\mu)}, \\ \mathbb{P}(\exists N \leq n \leq M, \widehat{\mu}_n \leq x) &\leq e^{-N(x-\mu)^2/2V}, \end{aligned} \tag{14}$$

and for every $x \geq \mu$,

$$\mathbb{P}(\exists N \leq n \leq M, \widehat{\mu}_n \geq x) \leq e^{-N \cdot \mathrm{kl}(x,\mu)}, \tag{15}$$

$$\mathbb{P}(\exists N \leq n \leq M, \widehat{\mu}_n \geq x) \leq e^{-N(x-\mu)^2/(2V)}. \tag{16}$$

This is the special case of the cited maximal inequality with its auxiliary tuning parameter fixed to one.

C Proofs of the Lower Bounds

C.1 Proof of Theorem 4.2

The proof of this theorem relies on the following lemma, whose proof is provided later in this section.

Lemma C.1. *Fix $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})$ and set*

$$g_i(\boldsymbol{\lambda}) := \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i}, \qquad i = 2, \dots, K.$$

Let $\mathcal{P}_{\le K}(\mathrm{Alt}_j(\mu_1))$ denote the set of probability measures on $\mathrm{Alt}_j(\mu_1)$ supported on at most K points. Then

$$\sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}_j(\mu_1)} \sum_{i=2}^{K} w_i g_i(\boldsymbol{\lambda}) = \inf_{P \in \mathcal{P}_{\le K}(\mathrm{Alt}_j(\mu_1))} \max_{i=2,\dots,K} \mathbb{E}_{\boldsymbol{\lambda} \sim P}\big[g_i(\boldsymbol{\lambda})\big]. \tag{17}$$

Our proof also requires the following concentration bound.

Lemma C.2 (Stopped exponential-family martingale bound with signed coefficients)**.** *Let $\mathcal{F} = (\mathcal{F}_t)_{t \ge 0}$ be a filtration. Let $(X_t)_{t \ge 1}$ be adapted to $\mathcal{F}$ and assume that, conditionally on $\mathcal{F}_{t-1}$, X_t follows the canonical one-dimensional exponential-family distribution ν_{θ_t}, where θ_t is $\mathcal{F}_{t-1}$-measurable. Let*

$$\mu_t := b'(\theta_t), \qquad \xi_t := X_t - \mu_t.$$

Then $(\xi_t)_{t \ge 1}$ is a martingale difference sequence.

Let $(x_t)_{t \ge 1}$ be a predictable real-valued sequence, i.e., x_t is $\mathcal{F}_{t-1}$-measurable and may be positive or negative. Define

$$S_t := \sum_{s=1}^{t} x_s \xi_s, \qquad W_t := \sum_{s=1}^{t} x_s^2.$$

Fix $0 < V < \infty$. A number $\lambda > 0$ is called admissible *if, for every $t \ge 1$ and $u \in [0,1]$,*

$$\theta_t + u\lambda x_t \in \Theta, \qquad b''(\theta_t + u\lambda x_t) \le V \quad \text{a.s.}$$

Then, for every admissible $\lambda > 0$, every almost surely finite stopping time τ, every $\beta > 0$, and every $v > 0$,

$$\mathbb{P}(S_\tau \ge \beta,\ W_\tau \le v) \le \exp\!\left(-\lambda\beta + \frac{\lambda^2 V v}{2}\right).$$

Consequently, if $\lambda^\star = \beta/(Vv)$ is admissible, then

$$\mathbb{P}(S_\tau \ge \beta,\ W_\tau \le v) \le \exp\!\left(-\frac{\beta^2}{2Vv}\right).$$

Proposition C.3. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \ge 0})$ be a filtered measurable space, and let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures. Let τ be a (possibly unbounded) stopping time and define the log-likelihood ratio*

$$L_\tau := \log \frac{d\mathbb{P}}{d\mathbb{Q}}\Big|_{\mathcal{F}_\tau}.$$

Then for any $\mathcal{F}_\tau$-measurable event A and any $\gamma \in \mathbb{R}$,

$$\mathbb{P}(A) \ \le\ e^{\gamma}\, \mathbb{Q}(A) \ +\ \mathbb{P}(A \cap \{L_\tau > \gamma\}). \tag{18}$$

Now, we are ready to prove Theorem 4.2.

Proof. Fix an instance $\boldsymbol{\mu}$ and $\xi > 0$. For each $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})$, Lemma C.1 gives a mixture $P^\star$ supported on at most K alternatives $\{\boldsymbol{\lambda}^k\}_{k=1}^K \subset \mathrm{Alt}_j(\mu_1)$, with weights $(q_k)_{k=1}^K$, such that

$$\max_{i \ge 2} \sum_{k=1}^{K} q_k \frac{\mathrm{kl}(\mu_i, \lambda_i^k)}{\Delta_i} \le \frac{1}{C_j(\boldsymbol{\mu})} + \xi.$$

Set

$$C_{j,\xi}(\boldsymbol{\mu}) := \big(C_j(\boldsymbol{\mu})^{-1} + \xi\big)^{-1}, \qquad C_\xi(\boldsymbol{\mu}) := \min_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_{j,\xi}(\boldsymbol{\mu}).$$

Fix these mixtures before letting $\delta \downarrow 0$. Define, for each $i = 2, \dots, K$,

$$c_i := \mathbb{E}_{\boldsymbol{\lambda} \sim P^\star}\left[\frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i}\right] = \sum_{k=1}^{K} q_k \frac{\mathrm{kl}(\mu_i, \lambda_i^k)}{\Delta_i}.$$

By the choice of the witness,

$$\max_{i=2,\dots,K} c_i \le \frac{1}{C_{j,\xi}(\boldsymbol{\mu})}.$$

Consequently, for any $\boldsymbol{w} = (w_2, \dots, w_K) \in \mathcal{W}$,

$$\sum_{i=2}^{K} w_i\, c_i \le \max_{i=2,\dots,K} c_i \le \frac{1}{C_{j,\xi}(\boldsymbol{\mu})}.$$

Applying the above inequality *pathwise* as follows. On the event $\{R_{\boldsymbol{\mu}}(\tau_\delta) > 0\}$, define the weights indexed by ω in the sample space,

$$w_i(\omega) \coloneqq \frac{\Delta_i N_i[\tau_\delta](\omega)}{R_{\boldsymbol{\mu}}(\tau_\delta)(\omega)}, \qquad i = 2, \dots, K.$$

Then for every ω with $R_{\boldsymbol{\mu}}(\tau_\delta)(\omega) > 0$, we have $\boldsymbol{w}(\omega) = (w_2(\omega), \dots, w_K(\omega)) \in \mathcal{W}$ and hence

$$\sum_{i=2}^{K} w_i(\omega) c_i \le \max_{i=2,\dots,K} c_i \le \frac{1}{C_{j,\xi}(\boldsymbol{\mu})}.$$

Multiplying by $R_{\boldsymbol{\mu}}(\tau_\delta)(\omega)$ yields, for all such ω,

$$\sum_{i=2}^{K} N_i[\tau_\delta](\omega) \mathbb{E}_{\boldsymbol{\lambda} \sim P^\star}\big[\mathrm{kl}(\mu_i, \lambda_i)\big] \le \frac{R_{\boldsymbol{\mu}}(\tau_\delta)(\omega)}{C_{j,\xi}(\boldsymbol{\mu})}.$$

On the event $\{R_{\boldsymbol{\mu}}(\tau_\delta) = 0\}$, the left-hand side is zero and the inequality holds trivially. Therefore,

$$\sum_{i=2}^{K} N_i[\tau_\delta] \mathbb{E}_{\boldsymbol{\lambda} \sim P^\star}\big[\mathrm{kl}(\mu_i, \lambda_i)\big] \le \frac{R_{\boldsymbol{\mu}}(\tau_\delta)}{C_{j,\xi}(\boldsymbol{\mu})}.$$

Equivalently, using the discrete representation of $P^\star$,

$$\sum_{k=1}^{K} q_k \sum_{i=2}^{K} N_i[\tau_\delta] \mathrm{kl}(\mu_i, \lambda_i^k) \le \frac{R_{\boldsymbol{\mu}}(\tau_\delta)}{C_{j,\xi}(\boldsymbol{\mu})} \tag{19}$$

Let $\mathbb{P}_{\boldsymbol{q}} = \sum_k q_k \mathbb{P}_{\boldsymbol{\lambda}^k}$ and $L = -\log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{q}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}\big|_{\mathcal{F}_{\tau_\delta}}$. Write ν_μ for the reward distribution with mean μ, and let $\theta(\mu)$ denote the unique parameter θ satisfying $b'(\theta) = \mu$. We introduce the empirical KL-divergence, which is defined as $\widehat{\mathrm{KL}}_t(\nu, \nu') = \frac{1}{t}\sum_{s=1}^{t} \log\left(\frac{d\nu}{d\nu'}(r_s)\right)$, where r_s denotes the s-th sample drawn from ν. Since $\log\big(p_\theta(x)/p_{\theta'}(x)\big) = (\theta - \theta')x - (b(\theta) - b(\theta'))$,

$$\widehat{\mathrm{KL}}_t(\nu_\theta, \nu_{\theta'}) = \mathrm{KL}(\nu_\theta, \nu_{\theta'}) + (\theta - \theta')(\widehat{\mu}_t - \mu). \tag{20}$$

Using convexity of $-\log$ and the Exponential Family likelihood decomposition (Eq. (20)) grouped by arms,

$$L \le \sum_{k=1}^{K} q_k \sum_{i=1}^{K} N_i[\tau_\delta] \mathrm{kl}(\mu_i, \lambda_i^k) + \sum_{k=1}^{K} q_k M_{\tau_\delta}(\boldsymbol{\mu}, \boldsymbol{\lambda}^k), \tag{21}$$

where

$$M_t(\boldsymbol{\mu}, \boldsymbol{\lambda}^k) = \sum_{i=1}^{K} \sum_{\ell=1}^{N_i(t)} (\theta(\mu_i) - \theta(\lambda_i^k))(r_{i,\ell} - \mu_i).$$

Here, $r_{i,\ell}$ denotes the reward of arm i on its ℓ-th pull. Since $\lambda_1^k = \mu_1$ for all k, the arm–1 martingale coefficient is zero, and hence

$$\sum_{k=1}^{K} q_k M_t(\boldsymbol{\mu}, \boldsymbol{\lambda}^k) = \sum_{i=2}^{K} \sum_{\ell=1}^{N_i(t)} \alpha_i (r_{i,\ell} - \mu_i), \qquad \alpha_i \coloneqq \theta(\mu_i) - \sum_{k=1}^{K} q_k \theta(\lambda_i^k).$$

Define

$$Q_\alpha := \max_{i\geq 2} \frac{\alpha_i^2}{\Delta_i}. \tag{22}$$

Let

$$S_t := \sum_{k=1}^{K} q_k M_t(\boldsymbol{\mu}, \boldsymbol{\lambda}^k) = \sum_{i=2}^{K} \sum_{\ell=1}^{N_i(t)} \alpha_i (r_{i,\ell} - \mu_i).$$

To apply Lemma C.2, write the same process in the chronological filtration. If the policy uses external randomization, we enlarge the filtration so that A_s is revealed before r_s is observed. Set $\alpha_1 = 0$. Set

$$x_s := \alpha_{A_s} \mathbb{1}\{A_s \neq 1\}, \qquad X_s := r_s, \qquad \theta_s := \theta(\mu_{A_s}), \qquad \mu_s := \mu_{A_s}.$$

Then x_s and θ_s are predictable, the coefficient x_s may be signed, and

$$S_t = \sum_{s=1}^{t} x_s (X_s - \mu_s), \qquad W_t := \sum_{s=1}^{t} x_s^2 = \sum_{i=2}^{K} \alpha_i^2 N_i(t).$$

On the event $\{R_{\boldsymbol{\mu}}(\tau_\delta) \leq r\}$,

$$W_{\tau_\delta} = \sum_{i=2}^{K} \alpha_i^2 N_i[\tau_\delta] \leq \sum_{i=2}^{K} Q_\alpha \Delta_i N_i[\tau_\delta] = Q_\alpha R_{\boldsymbol{\mu}}(\tau_\delta) \leq Q_\alpha r.$$

Therefore,

$$\{R_{\boldsymbol{\mu}}(\tau_\delta) \leq r\} \subseteq \{W_{\tau_\delta} \leq Q_\alpha r\}.$$

Since the true natural parameters are interior points of Θ and there are finitely many coefficients α_i, there exist $\lambda_0 > 0$ and $V_{\mathrm{loc}} < \infty$ such that

$$\theta(\mu_i) + u\lambda\alpha_i \in \Theta, \quad b''(\theta(\mu_i) + u\lambda\alpha_i) \leq V_{\mathrm{loc}} \quad (0 \leq \lambda \leq \lambda_0,\ 0 \leq u \leq 1).$$

There are finitely many valid answers j, so choose common constants for all their witnesses and take Q_α to be the largest of their values. If $Q_\alpha = 0$, then $S_t = 0$ and its upper-tail probability is zero. Otherwise, for $\beta/(V_{\mathrm{loc}} Q_\alpha r) \leq \lambda_0$, applying Lemma C.2 with $v = Q_\alpha r$ and $\lambda = \beta/(V_{\mathrm{loc}} Q_\alpha r)$ gives

$$\mathbb{P}_{\boldsymbol{\mu}}(S_{\tau_\delta} > \beta,\ R_{\boldsymbol{\mu}}(\tau_\delta) \leq r) \leq \mathbb{P}_{\boldsymbol{\mu}}(S_{\tau_\delta} > \beta,\ W_{\tau_\delta} \leq Q_\alpha r) \tag{23}$$

$$\leq \exp\left(-\frac{\beta^2}{2V_{\mathrm{loc}} Q_\alpha r}\right). \tag{24}$$

By (19),

$$\sum_{k=1}^{K} q_k \sum_{i=2}^{K} N_i[\tau_\delta] \,\mathrm{kl}(\mu_i, \lambda_i^k) \ \leq \ \frac{R_{\boldsymbol{\mu}}(\tau_\delta)}{C_{j,\xi}(\boldsymbol{\mu})}.$$

Hence on $\{R_{\boldsymbol{\mu}}(\tau_\delta) \leq r\}$,

$$\sum_{k=1}^{K} q_k \sum_{i=2}^{K} N_i[\tau_\delta] \,\mathrm{kl}(\mu_i, \lambda_i^k) \ \leq \ \frac{r}{C_{j,\xi}(\boldsymbol{\mu})}. \tag{25}$$

Thus,

$$L \ \leq \ \frac{r}{C_{j,\xi}(\boldsymbol{\mu})} \ + \ S_{\tau_\delta} \quad \text{on } \{R_{\boldsymbol{\mu}}(\tau_\delta) \leq r\}.$$

For any $\mathcal{F}_{\tau_\delta}$-measurable $A \subseteq \{R_{\boldsymbol{\mu}}(\tau_\delta) \leq r\}$ and any $\beta > 0$ for which the above martingale bound applies, using Proposition C.3 with $\gamma = \frac{r}{C_{j,\xi}(\boldsymbol{\mu})} + \beta$, we have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu}}(A) &\leq e^{\gamma} \sum_{k=1}^{K} q_k \mathbb{P}_{\boldsymbol{\lambda}^k}(A) + \mathbb{P}_{\boldsymbol{\mu}}(A \cap \{L > \gamma\}) \\
&\leq e^{\gamma} \sum_{k=1}^{K} q_k \mathbb{P}_{\boldsymbol{\lambda}^k}(A) + \mathbb{P}_{\boldsymbol{\mu}}\big(S_{\tau_\delta} > \beta, R_{\boldsymbol{\mu}}(\tau_\delta) \leq r\big) \\
&\leq \exp\left(\frac{r}{C_{j,\xi}(\boldsymbol{\mu})} + \beta\right) \sum_{k=1}^{K} q_k \mathbb{P}_{\boldsymbol{\lambda}^k}(A) + \exp\left(-\frac{\beta^2}{2V_{\mathrm{loc}} Q_\alpha r}\right)
\end{aligned} \tag{26}$$

where the last inequality is due to (23). Since the policy is (ε, δ)-PAC, so

$$\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r\big) \le \delta + \sum_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} \mathbb{P}_{\boldsymbol{\mu}}(R_{\boldsymbol{\mu}}(\tau_\delta) \le r,\ i_{\text{out}} = j). \tag{27}$$

For each fixed $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})$, apply (26) with $A = \{R_{\boldsymbol{\mu}}(\tau_\delta) \le r,\ i_{\text{out}} = j\}$ to get

$$\mathbb{P}_{\boldsymbol{\mu}}(A) \le \exp\!\left(\frac{r}{C_{j,\xi}(\boldsymbol{\mu})} + \beta\right)\delta \ + \ \exp\!\left(-\frac{\beta^2}{2V_{\text{loc}}Q_\alpha r}\right).$$

Substituting this bound into (27) yields

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r\big) &\le \delta + \sum_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} \left[\exp\!\left(\frac{r}{C_{j,\xi}(\boldsymbol{\mu})} + \beta\right)\delta + \exp\!\left(-\frac{\beta^2}{2V_{\text{loc}}Q_\alpha r}\right)\right] \\ &= \delta + \delta \sum_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} \exp\!\left(\frac{r}{C_{j,\xi}(\boldsymbol{\mu})} + \beta\right) + |\mathcal{I}_\varepsilon(\boldsymbol{\mu})| \exp\!\left(-\frac{\beta^2}{2V_{\text{loc}}Q_\alpha r}\right).\end{aligned}$$

Let $c_1 = |\mathcal{I}_\varepsilon(\boldsymbol{\mu})| \le K$. Moreover, since $C_\xi(\boldsymbol{\mu}) := \min_{i \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_{i,\xi}(\boldsymbol{\mu})$, we have $C_{j,\xi}(\boldsymbol{\mu}) \ge C_\xi(\boldsymbol{\mu})$ for all $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})$, and hence

$$\exp\!\left(\frac{r}{C_{j,\xi}(\boldsymbol{\mu})} + \beta\right) \le \exp\!\left(\frac{r}{C_\xi(\boldsymbol{\mu})} + \beta\right).$$

Therefore,

$$\sum_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} \exp\!\left(\frac{r}{C_{j,\xi}(\boldsymbol{\mu})} + \beta\right) \le c_1 \exp\!\left(\frac{r}{C_\xi(\boldsymbol{\mu})} + \beta\right),$$

which gives

$$\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r\big) \le \delta + c_1 \exp\!\left(\frac{r}{C_\xi(\boldsymbol{\mu})} + \beta\right)\delta + c_1 \exp\!\left(-\frac{\beta^2}{2V_{\text{loc}}Q_\alpha r}\right).$$

Pick any $\eta \in (0,1)$ and set $r = (1-\eta)C_\xi(\boldsymbol{\mu})\log(1/\delta)$ and $\beta = \frac{\eta}{2}\log(1/\delta)$. For $Q_\alpha > 0$, the chosen tilt is

$$\lambda^\star = \frac{\eta}{2V_{\text{loc}}Q_\alpha(1-\eta)C_\xi(\boldsymbol{\mu})},$$

which is at most λ_0 for sufficiently small fixed $\eta > 0$. If $Q_\alpha = 0$, the martingale term is zero. Thus $\mathbb{P}_{\boldsymbol{\mu}}(R_{\boldsymbol{\mu}}(\tau_\delta) \le r) \to 0$ as $\delta \downarrow 0$. Note that for any $r > 0$, by the Markov inequality,

$$\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)] \ \ge \ r\,\mathbb{P}_{\boldsymbol{\mu}}(R_{\boldsymbol{\mu}}(\tau_\delta) \ge r) \ge r\big(1 - \mathbb{P}_{\boldsymbol{\mu}}(R_{\boldsymbol{\mu}}(\tau_\delta) \le r)\big).$$

Letting $\eta \downarrow 0$ and then $\xi \downarrow 0$ concludes the theorem, since $C_\xi(\boldsymbol{\mu}) \uparrow C(\boldsymbol{\mu})$. □

C.2 Proof of Theorem 4.5

Proof. In this subsection, we adopt the same notation used in the proof of Theorem 4.2. Let $j = 1$ and fix $\xi > 0$. Use the finite witness from the preceding proof and write $C^*_\xi(\boldsymbol{\mu}) = (C^*(\boldsymbol{\mu})^{-1} + \xi)^{-1}$. Fix $r > 0$ and $\beta > 0$ for which the preceding martingale bound applies. Recall $C^*_\xi(\boldsymbol{\mu}) = C_{1,\xi}(\boldsymbol{\mu})$. For the event $A = \{R_{\boldsymbol{\mu}}(\tau_\delta) \le r,\ i_{\text{out}} = 1\}$, similar to (26) in the proof of Theorem 4.2, we take $\gamma = \frac{r}{C^*_\xi(\boldsymbol{\mu})} + \beta$, which yields

$$\mathbb{P}_{\boldsymbol{\mu}}(A) \ \le \ \exp\!\left(\frac{r}{C^*_\xi(\boldsymbol{\mu})} + \beta\right)\sum_{k=1}^{K} q_k\,\mathbb{P}_{\boldsymbol{\lambda}^k}(A) \ + \ \mathbb{P}_{\boldsymbol{\mu}}\big(S_{\tau_\delta} > \beta,\ R_{\boldsymbol{\mu}}(\tau_\delta) \le r\big).$$

Under each $\boldsymbol{\lambda}^k \in \text{Alt}_1(\mu_1)$, the arm 1 is *not* an ε-best arm, hence the fact that the policy is (ε, δ)-PAC implies $\mathbb{P}_{\boldsymbol{\lambda}^k}(i_{\text{out}} = 1) \le \delta$. Therefore

$$\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r,\ i_{\text{out}} = 1\big) \ \le \ \exp\left(\frac{r}{C^*_\xi(\boldsymbol{\mu})} + \beta\right)\delta \ + \ \exp\left(-\frac{\beta^2}{2V_{\text{loc}}Q_\alpha r}\right), \tag{28}$$

where this derivation is similar to that leading to (26) and Q_α is defined in (22). Fix any $\eta \in (0,1)$ and set

$$r \;\coloneqq\; (1-\eta)C^*_\xi(\boldsymbol{\mu})\log\frac{1}{\delta}, \qquad \beta \;\coloneqq\; \frac{\eta}{2}\log\frac{1}{\delta}.$$

With the local constants from the preceding proof, the chosen tilt is $\lambda^\star = \eta/[2V_{\mathrm{loc}}Q_\alpha(1-\eta)C^*_\xi(\boldsymbol{\mu})]$. It is admissible for sufficiently small fixed $\eta > 0$; if $Q_\alpha = 0$, the martingale term is identically zero. Plugging these choices into (28) yields

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r,\; i_{\mathrm{out}} = 1\big) &\le \exp\Big(\frac{r}{C^*_\xi(\boldsymbol{\mu})} + \beta\Big)\,\delta \;+\; \exp\Big(-\frac{\beta^2}{2V_{\mathrm{loc}}Q_\alpha r}\Big) \\
&= \exp\Big((1-\eta)\log\frac{1}{\delta} + \frac{\eta}{2}\log\frac{1}{\delta}\Big)\,\delta \\
&\quad + \exp\Big(-\frac{(\eta^2/4)\log^2(1/\delta)}{2V_{\mathrm{loc}}Q_\alpha(1-\eta)C^*_\xi(\boldsymbol{\mu})\log(1/\delta)}\Big) \\
&= \delta^{\eta/2} \;+\; \delta^{\frac{\eta^2}{8V_{\mathrm{loc}}Q_\alpha(1-\eta)C^*_\xi(\boldsymbol{\mu})}}.
\end{aligned}$$

In particular, both exponents are strictly positive, hence

$$\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r,\; i_{\mathrm{out}} = 1\big) \xrightarrow[\delta\downarrow 0]{} 0.$$

Since the policy belongs to an ε-PAC* family, we have $\mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out}} = 1) \to 1$ as $\delta \downarrow 0$, equivalently $\mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out}} \neq 1) \to 0$. Therefore,

$$\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r\big) \le \mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r,\; i_{\mathrm{out}} = 1\big) + \mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out}} \neq 1) \xrightarrow[\delta\downarrow 0]{} 0,$$

which implies

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\frac{R_{\boldsymbol{\mu}}(\tau_\delta)}{\log(1/\delta)} \le (1-\eta)C^*_\xi(\boldsymbol{\mu})\right) = \mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r\big) \xrightarrow[\delta\downarrow 0]{} 0.$$

Finally, using $R_{\boldsymbol{\mu}}(\tau_\delta) \ge 0$ and $\mathbb{E}[X] \ge a\,\mathbb{P}(X \ge a)$ for any $X \ge 0$ and $a > 0$, we obtain

$$\frac{\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} \ge \frac{r}{\log(1/\delta)}\,\mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \ge r\big) \ge (1-\eta)C^*_\xi(\boldsymbol{\mu})\Big(1 - \mathbb{P}_{\boldsymbol{\mu}}\big(R_{\boldsymbol{\mu}}(\tau_\delta) \le r\big)\Big).$$

Taking $\liminf_{\delta\downarrow 0}$ and using $\mathbb{P}_{\boldsymbol{\mu}}(R_{\boldsymbol{\mu}}(\tau_\delta) \le r) \to 0$ gives

$$\liminf_{\delta\downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} \ge (1-\eta)C^*_\xi(\boldsymbol{\mu}).$$

Letting $\eta \downarrow 0$ gives , for each $\xi > 0$,

$$\liminf_{\delta\downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} \ge C^*_\xi(\boldsymbol{\mu}).$$

Letting $\xi \downarrow 0$ gives the claimed lower bound $C^*(\boldsymbol{\mu})$. □

C.3 Proof of Theorem 4.8

Proof. We follow the sample–regret trade-off argument of Yang et al. [31], using alternatives in which arm 1 is no longer ε-best. Fix an instance $\boldsymbol{\mu}$ with a unique best arm 1 and an ε-PAC* family $(\pi_\delta)_{\delta\in(0,1)}$ satisfying

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

for all sufficiently small δ. Write $L_\delta = \log(1/\delta)$ and $p_\delta = \mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out},\delta} = 1)$. By the ε-PAC* property, $p_\delta \to 1$ as $\delta \downarrow 0$.

Let $u = \mu_1 + \varepsilon$. For a sufficiently small, fixed $\rho > 0$, both $\mu_1 - \rho$ and $u - \rho/2$ belong to the mean-parameter space I, and $u - \rho/2 > \mu_a$ for every $a \ge 2$. For each $a \in \{2,\dots,K\}$, define an alternative $\boldsymbol{\lambda}^{a,\rho}$ by

$$\lambda^{a,\rho}_1 = \mu_1 - \rho, \qquad \lambda^{a,\rho}_a = u - \frac{\rho}{2}, \qquad \lambda^{a,\rho}_i = \mu_i \quad (i \notin \{1,a\}).$$

Since

$$\lambda_a^{a,\rho} - \lambda_1^{a,\rho} = \varepsilon + \frac{\rho}{2} > \varepsilon,$$

arm 1 is not ε-best under $\boldsymbol{\lambda}^{a,\rho}$. Thus the (ε, δ)-PAC property implies

$$q_{a,\delta} := \mathbb{P}_{\boldsymbol{\lambda}^{a,\rho}}(i_{\mathrm{out},\delta} = 1) \le \delta.$$

Set

$$d_1(\rho) = \mathrm{kl}(\mu_1, \mu_1 - \rho), \qquad D_a(\rho) = \mathrm{kl}\Big(\mu_a, u - \frac{\rho}{2}\Big).$$

Applying the stopped change-of-measure inequality (Lemma 1 of Kaufmann et al. [22]) to the event $\{i_{\mathrm{out},\delta} = 1\}$ and the two instances $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}^{a,\rho}$ gives

$$\begin{aligned} &d_1(\rho)\,\mathbb{E}_{\boldsymbol{\mu}}[N_1[\tau_\delta]] + D_a(\rho)\,\mathbb{E}_{\boldsymbol{\mu}}[N_a[\tau_\delta]] \\ &\quad \ge \mathrm{kl}_{\mathcal{B}}(p_\delta, q_{a,\delta}) \ge p_\delta L_\delta - \log 2, \end{aligned} \tag{29}$$

where the last inequality follows from $\mathrm{kl}_{\mathcal{B}}(p, q) \ge p \log(1/q) - \log 2$ and $q_{a,\delta} \le \delta$.

Multiplying (29) by $\Delta_a / D_a(\rho)$ and summing over $a \ge 2$ yields

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)] &= \sum_{a=2}^{K} \Delta_a\,\mathbb{E}_{\boldsymbol{\mu}}[N_a[\tau_\delta]] \\ &\ge C_\rho\Big(p_\delta L_\delta - \log 2 - d_1(\rho)\,\mathbb{E}_{\boldsymbol{\mu}}[N_1[\tau_\delta]]\Big), \end{aligned} \tag{30}$$

where

$$C_\rho := \sum_{a=2}^{K} \frac{\Delta_a}{D_a(\rho)}.$$

Since $\mathbb{E}_{\boldsymbol{\mu}}[N_1[\tau_\delta]] \le \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \le \kappa L_\delta$, dividing (30) by L_δ and letting $\delta \downarrow 0$ gives, for every sufficiently small fixed $\rho > 0$,

$$\liminf_{\delta \downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} \ge \big(1 - \kappa d_1(\rho)\big) C_\rho. \tag{31}$$

It remains to show that the right-hand side is strictly larger than $C^*(\boldsymbol{\mu})$ for some fixed $\rho > 0$. Recall that

$$C^*(\boldsymbol{\mu}) = \sum_{a=2}^{K} \frac{\Delta_a}{\mathrm{kl}(\mu_a, u)}.$$

For a one-parameter exponential family,

$$\frac{\partial}{\partial y} \mathrm{kl}(\mu_a, y) = \frac{y - \mu_a}{V(y)}.$$

Hence, as $\rho \downarrow 0$,

$$D_a(\rho) = \mathrm{kl}(\mu_a, u) - \frac{\rho}{2} \frac{u - \mu_a}{V(u)} + o(\rho),$$

and therefore

$$C_\rho = C^*(\boldsymbol{\mu}) + c_{\boldsymbol{\mu},\varepsilon}\rho + o(\rho), \tag{32}$$

where

$$c_{\boldsymbol{\mu},\varepsilon} := \frac{1}{2V(u)} \sum_{a=2}^{K} \frac{\Delta_a (u - \mu_a)}{\mathrm{kl}(\mu_a, u)^2} > 0.$$

On the other hand, the local quadratic expansion of the KL divergence gives

$$d_1(\rho) = \frac{\rho^2}{2V(\mu_1)} + o(\rho^2).$$

Combining these expansions, for the fixed constant κ,

$$\big(1 - \kappa d_1(\rho)\big) C_\rho = C^*(\boldsymbol{\mu}) + c_{\boldsymbol{\mu},\varepsilon}\rho + o(\rho).$$

Since $c_{\boldsymbol{\mu},\varepsilon} > 0$, we can choose a sufficiently small *fixed* $\rho_* > 0$, independent of δ, such that

$$\big(1 - \kappa d_1(\rho_*)\big) C_{\rho_*} > C^*(\boldsymbol{\mu}).$$

Applying (31) with this ρ_* proves

$$\liminf_{\delta \downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} > C^*(\boldsymbol{\mu}).$$

□

C.4 Proof of Supporting Lemmas in the Lower Bound Part

Proof of Lemma C.1. The proof is the regret-normalized analogue of the minimax witness lemma of Degenne and Koolen [10]. We give the argument explicitly in our notation in order to avoid any compactness or attainment assumption.

Let

$$\mathcal{S}_j := \left\{ \big(g_2(\boldsymbol{\lambda}), \dots, g_K(\boldsymbol{\lambda})\big) : \boldsymbol{\lambda} \in \mathrm{Alt}_j(\mu_1) \right\} \subseteq \mathbb{R}_+^{K-1}.$$

For $s = (s_2, \dots, s_K) \in \mathbb{R}^{K-1}$, define

$$h(s) := \max_{i=2,\dots,K} s_i.$$

Since a mixed strategy of MAX is exactly an element of the regret simplex $\mathcal{W}$, after dropping the zero first coordinate, we have

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}_j(\mu_1)} \sum_{i=2}^{K} w_i g_i(\boldsymbol{\lambda}) = \sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{s\in\mathcal{S}_j} \sum_{i=2}^{K} w_i s_i.$$

Moreover, because the map $s \mapsto \sum_{i=2}^{K} w_i s_i$ is linear, the infimum over $\mathcal{S}_j$ is the same as the infimum over its convex hull:

$$\inf_{s\in\mathcal{S}_j} \sum_{i=2}^{K} w_i s_i = \inf_{s\in\mathrm{conv}(\mathcal{S}_j)} \sum_{i=2}^{K} w_i s_i.$$

It remains to show that

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{s\in\mathrm{conv}(\mathcal{S}_j)} \sum_{i=2}^{K} w_i s_i = \inf_{s\in\mathrm{conv}(\mathcal{S}_j)} h(s).$$

Let

$$m := \inf_{s\in\mathrm{conv}(\mathcal{S}_j)} h(s).$$

First, for every $\boldsymbol{w} \in \mathcal{W}$ and every $s \in \mathrm{conv}(\mathcal{S}_j)$,

$$\sum_{i=2}^{K} w_i s_i \le h(s),$$

because $\boldsymbol{w}$ is a probability vector over $\{2, \dots, K\}$. Hence

$$\inf_{s\in\mathrm{conv}(\mathcal{S}_j)} \sum_{i=2}^{K} w_i s_i \le \inf_{s\in\mathrm{conv}(\mathcal{S}_j)} h(s) = m.$$

Taking the supremum over $\boldsymbol{w} \in \mathcal{W}$ gives

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \inf_{s\in\mathrm{conv}(\mathcal{S}_j)} \sum_{i=2}^{K} w_i s_i \le m.$$

We now prove the reverse inequality. Fix any $r < m$. Define the open convex set

$$\mathcal{O}_r := \{ s \in \mathbb{R}^{K-1} : h(s) < r \} = (-\infty, r)^{K-1}.$$

By the definition of m, the sets $\mathrm{conv}(\mathcal{S}_j)$ and $\mathcal{O}_r$ are disjoint. By the separating hyperplane theorem, there exists a nonzero vector $y = (y_2, \dots, y_K) \in \mathbb{R}^{K-1}$ such that

$$\sum_{i=2}^{K} y_i s_i \ge \sum_{i=2}^{K} y_i u_i, \qquad \forall s \in \mathrm{conv}(\mathcal{S}_j),\ \forall u \in \mathcal{O}_r.$$

Since $\mathcal{O}_r = (-\infty, r)^{K-1}$ is unbounded below in each coordinate, the above separation is possible only if $y_i \ge 0$ for every $i = 2, \dots, K$. Since $y \ne 0$, we have $\sum_{i=2}^{K} y_i > 0$. Moreover,

$$\sup_{u\in\mathcal{O}_r} \sum_{i=2}^{K} y_i u_i = r \sum_{i=2}^{K} y_i.$$

Therefore, for every $s \in \operatorname{conv}(\mathcal{S}_j)$,

$$\sum_{i=2}^{K} y_i s_i \geq r \sum_{i=2}^{K} y_i.$$

Define

$$w_i := \frac{y_i}{\sum_{\ell=2}^{K} y_\ell}, \qquad i = 2, \ldots, K,$$

and set $w_1 = 0$. Then $\boldsymbol{w} \in \mathcal{W}$, and

$$\inf_{s \in \operatorname{conv}(\mathcal{S}_j)} \sum_{i=2}^{K} w_i s_i \geq r.$$

Consequently,

$$\sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{s \in \operatorname{conv}(\mathcal{S}_j)} \sum_{i=2}^{K} w_i s_i \geq r.$$

Letting $r \uparrow m$ yields the reverse inequality. Hence

$$\sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{s \in \mathcal{S}_j} \sum_{i=2}^{K} w_i s_i = \inf_{s \in \operatorname{conv}(\mathcal{S}_j)} \max_{i=2,\ldots,K} s_i.$$

Finally, by Carathéodory's theorem in dimension $K-1$, every point $s \in \operatorname{conv}(\mathcal{S}_j)$ can be written as a convex combination of at most K points of $\mathcal{S}_j$. Therefore,

$$\inf_{s \in \operatorname{conv}(\mathcal{S}_j)} \max_{i=2,\ldots,K} s_i = \inf_{P \in \mathcal{P}_{\leq K}(\operatorname{Alt}_j(\mu_1))} \max_{i=2,\ldots,K} \mathbb{E}_{\boldsymbol{\lambda} \sim P}\big[g_i(\boldsymbol{\lambda})\big],$$

which proves (17). □

Proof of Lemma C.2. Since θ_t is $\mathcal{F}_{t-1}$-measurable and $X_t \mid \mathcal{F}_{t-1} \sim \nu_{\theta_t}$,

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[X_t - \mu_t \mid \mathcal{F}_{t-1}] = b'(\theta_t) - \mu_t = 0.$$

Thus $(\xi_t)_{t \geq 1}$ is a martingale difference sequence.

Fix $\lambda > 0$ satisfying the parameter-segment and variance conditions in Lemma C.2, and define

$$M_t := \exp\left(\lambda S_t - \frac{\lambda^2 V}{2} W_t\right), \qquad t \geq 0,$$

with $S_0 = W_0 = 0$. We prove that $(M_t)_{t \geq 0}$ is a nonnegative supermartingale. Since x_t is $\mathcal{F}_{t-1}$-measurable,

$$S_t - S_{t-1} = x_t \xi_t, \qquad W_t - W_{t-1} = x_t^2.$$

Therefore,

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1} \exp\left(-\frac{\lambda^2 V}{2} x_t^2\right) \mathbb{E}[\exp(\lambda x_t \xi_t) | \mathcal{F}_{t-1}]\,.$$

Let $h_t := \lambda x_t$. Notice that h_t may be positive, zero, or negative. The exponential-family moment-generating identity gives

$$\mathbb{E}[\exp(h_t(X_t - \mu_t)) | \mathcal{F}_{t-1}] = \exp(b(\theta_t + h_t) - b(\theta_t) - h_t b'(\theta_t))\,.$$

By Taylor's formula with integral remainder, for any signed h_t such that $\{\theta_t + u h_t : u \in [0,1]\} \subseteq \Theta$,

$$b(\theta_t + h_t) - b(\theta_t) - h_t b'(\theta_t) = h_t^2 \int_0^1 (1-u) b''(\theta_t + u h_t)\, \mathrm{d}u.$$

The assumed variance bound on the tilted segment yields

$$b(\theta_t + h_t) - b(\theta_t) - h_t b'(\theta_t) \leq h_t^2 \int_0^1 (1-u) V\, \mathrm{d}u = \frac{V h_t^2}{2} = \frac{\lambda^2 V x_t^2}{2}.$$

Hence

$$\mathbb{E}[\exp(\lambda x_t \xi_t)|\mathcal{F}_{t-1}] \le \exp\left(\frac{\lambda^2 V x_t^2}{2}\right),$$

and consequently $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \le M_{t-1}$. Thus $(M_t)_{t\ge 0}$ is a nonnegative supermartingale.

Let $\tau_n := \tau \wedge n$. Since τ_n is bounded, optional stopping for nonnegative supermartingales gives

$$\mathbb{E}[M_{\tau_n}] \le \mathbb{E}[M_0] = 1.$$

Since $\tau < \infty$ almost surely, $M_{\tau_n} \to M_\tau$ almost surely. By Fatou's lemma,

$$\mathbb{E}[M_\tau] = \mathbb{E}\Big[\liminf_{n\to\infty} M_{\tau_n}\Big] \le \liminf_{n\to\infty} \mathbb{E}[M_{\tau_n}] \le 1.$$

On the event $\{S_\tau \ge \beta,\ W_\tau \le v\}$,

$$M_\tau \ge \exp\left(\lambda\beta - \frac{\lambda^2 V v}{2}\right).$$

Therefore, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}(S_\tau \ge \beta,\ W_\tau \le v) &\le \mathbb{P}\left(M_\tau \ge \exp\left(\lambda\beta - \frac{\lambda^2 V v}{2}\right)\right) \\
&\le \exp\left(-\lambda\beta + \frac{\lambda^2 V v}{2}\right)\mathbb{E}[M_\tau] \\
&\le \exp\left(-\lambda\beta + \frac{\lambda^2 V v}{2}\right).
\end{aligned}$$

This proves the first claim. If $\lambda^\star = \beta/(Vv)$ is admissible, substituting this choice gives

$$-\lambda^\star\beta + \frac{(\lambda^\star)^2 V v}{2} = -\frac{\beta^2}{2Vv},$$

which proves the optimized bound. □

Proof of Proposition C.3. By the Radon–Nikodym theorem on $\mathcal{F}_\tau$,

$$\mathbb{Q}(A) = \mathbb{E}_{\mathbb{Q}}[\mathbb{1}[A]] = \mathbb{E}_{\mathbb{P}}[\mathbb{1}[A]\exp(-L_\tau)].$$

Therefore,

$$\begin{aligned}
\mathbb{Q}(A) &\ge \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\{A \cap \{L_\tau \le \gamma\}\}\exp(-L_\tau)\right] \\
&\ge e^{-\gamma}\mathbb{P}\left(A \cap \{L_\tau \le \gamma\}\right) \\
&\ge e^{-\gamma}\big(\mathbb{P}(A) - \mathbb{P}(A \cap \{L_\tau > \gamma\})\big).
\end{aligned}$$

Rearranging gives

$$\mathbb{P}(A) \le e^{\gamma}\mathbb{Q}(A) + \mathbb{P}(A \cap \{L_\tau > \gamma\}),$$

as desired. □

D Proof of Correctness, Regret Bound and Sample Complexity of `KLUCB+Greedy`

D.1 Proof of the Asymptotic Regret Bound ((12) in Theorem 5.1)

Throughout this subsection, δ is sufficiently small that $\bar{\delta} = \delta$. In the asymptotic analysis, we use $a_\delta \lesssim b_\delta$ to denote that there exists some $\delta_0 \in (0,1)$, such that for all $0 \le \delta \le \delta_0$, $a_\delta \le b_\delta$. Similarly, we use $a_\delta \gtrsim b_\delta$ to denote that there exists some $\delta_0 \in (0,1)$, such that for all $0 \le \delta \le \delta_0$, $a_\delta \ge b_\delta$. Moreover, recall that $f(\delta) = O(g(\delta))$ as $\delta \downarrow 0$ means that there exist constants $C > 0$ and $\delta_0 \in (0,1)$ such that for all $\delta \in (0, \delta_0)$, $|f(\delta)| \le C\,|g(\delta)|$. Equivalently, $\limsup_{\delta\downarrow 0}\left|\frac{f(\delta)}{g(\delta)}\right| < \infty$.

Fix a suboptimal arm $i \neq 1$. Let

$$\zeta_1 := \frac{1}{\sqrt{\log\log(1/\delta)}}, \qquad \theta_\delta := \mu_1 + \varepsilon - 5\zeta_1.$$

For all sufficiently small δ, we have $\zeta = \zeta_1$ in Algorithm 1. In the rest of this regret proof, we therefore write the algorithmic slack as ζ_1. Moreover, for all sufficiently small δ, $\zeta_1 \leq \varepsilon/10$, $\theta_\delta \in (\mu_1, \mu_1 + \varepsilon)$, $\Delta_i - 5\zeta_1/2 > 0$, and $\theta_\delta > \mu_i + \zeta_1$.

Let B_t denote the event that Algorithm 1 uses the UCB branch at time t, so that $A_{t+1} = b_t$. Let G_t denote the event that the algorithm uses the candidate/greedy branch at time t, so that $A_{t+1} = \mathrm{a}^*(t)$. Here t is the number of observations before pulling A_{t+1}. Both branch events are false for $t < K$ and after stopping. During the initialization rounds, set both B_t and G_t to be false.

We further decompose the number of pulls of arm i by the value $U_{b_t}(t)$. Since each arm is pulled once during initialization, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}[N_i[\tau_\delta]] \leq 1 &+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) \geq \theta_\delta\}\right]}_{I_1} \\
&+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta\}\right]}_{I_2} \\
&+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{B}_t\}\right]}_{I_3}. \qquad (33)
\end{aligned}$$

Bounding Term I_1. Assume $U_{b_t}(t) \geq \theta_\delta$, $A_{t+1} = i$, and G_t holds. Then if $i \in \mathcal{C}_t$,

$$\begin{aligned}
\widehat{\mu}_i(t) &\geq U_{b_t}(t) - \varepsilon + \frac{7\zeta_1}{2} \\
&\geq \theta_\delta - \varepsilon + \frac{7\zeta_1}{2} = \mu_1 - \frac{3\zeta_1}{2}.
\end{aligned}$$

Otherwise $i \notin \mathcal{C}_t$. Since G_t holds, we must have $\mathcal{C}_t \neq \emptyset$, $\mathrm{a}^*(t) = \mathrm{a}^*(t-1)$, and $i = \mathrm{a}^*(t)$. Moreover, the greedy update condition in Line 13 of Algorithm 1 is not satisfied. Hence

$$\begin{aligned}
& \max_{j \in \mathcal{C}_t} \widehat{\mu}_j(t) < \widehat{\mu}_i(t) + \zeta_1 && \text{(According to Line 13 of Algorithm 1)} \\
\Longrightarrow \quad & \widehat{\mu}_i(t) \geq \max_{j \in \mathcal{C}_t} \widehat{\mu}_j(t) - \zeta_1 \\
& \geq U_{b_t}(t) - \varepsilon + \frac{7\zeta_1}{2} - \zeta_1 \\
& \geq \theta_\delta - \varepsilon + \frac{5\zeta_1}{2} = \mu_1 - \frac{5\zeta_1}{2}.
\end{aligned}$$

Therefore,

$$I_1 \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{A_{t+1} = i,\, \widehat{\mu}_i(t) \geq \mu_1 - \frac{5\zeta_1}{2}\right\}\right].$$

Let $\widehat{\mu}_{i,s} = \widehat{\mu}_i(t)$ with $s = N_i(t)$. Thanks to Lemma B.4, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{A_{t+1} = i,\, \widehat{\mu}_i(t) \geq \mu_1 - \frac{5\zeta_1}{2}\right\}\right] \\
&\quad\leq \sum_{s=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i,s} \geq \mu_1 - \frac{5\zeta_1}{2}\right) \\
&\quad\leq \sum_{s=1}^{T} \exp\left(-\frac{s(\Delta_i - 5\zeta_1/2)^2}{2V}\right) \\
&\quad\leq \sum_{s=1}^{\infty} \exp\left(-\frac{s(\Delta_i - 5\zeta_1/2)^2}{2V}\right) = O(1), \qquad (34)
\end{aligned}$$

where the last equality follows from $\zeta_1 \to 0$ and $\Delta_i > 0$. Therefore,

$$I_1 = O(1).$$

Bounding Term I_2. Let $U_{i,s} = U_i(t)$ for $N_i(t) = s$. Let $M = \lceil \log(1/\delta)/(\log\log(1/\delta))^4 \rceil$ and let x_s satisfy $x_s \leq \theta_\delta$ and $\mathrm{kl}(x_s, \theta_\delta) = f_s(1)/s$. We have

$$\begin{aligned}
\mathbb{P}\Big(\exists t \leq T:\, N_1(t) \leq M,\, U_1(t) \leq \theta_\delta\Big) &\leq \mathbb{P}\Big(\exists s \leq M:\, U_{1,s} \leq \theta_\delta\Big) \\
&\leq \sum_{s=1}^{M} \mathbb{P}\Big(U_{1,s} \leq \theta_\delta\Big) \\
&\leq \sum_{s=1}^{M} \mathbb{P}\Big(\widehat{\mu}_{1,s} \leq \theta_\delta,\, \mathrm{kl}(\widehat{\mu}_{1,s}, \theta_\delta) > \frac{f_s(1)}{s}\Big). \qquad (35)
\end{aligned}$$

Define $V_{\min} = \min_{\mu \in (\mu_1, \mu_1+\varepsilon)} V(\mu)$. Note that

$$\begin{aligned}
\mathrm{kl}(\widehat{\mu}_{1,s}, \theta_\delta) &= \mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) + \int_{\mu_1}^{\theta_\delta} \frac{x - \widehat{\mu}_{1,s}}{V(x)} \mathrm{d}x \\
&\leq \mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) + \varepsilon \cdot (\mu_1 + \varepsilon - \widehat{\mu}_{1,s})/V_{\min} \\
&\leq \mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) + \varepsilon \cdot \left(\varepsilon + \sqrt{2V\mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1)}\right)/V_{\min}, \qquad (36)
\end{aligned}$$

where the last inequality is due to Lemma B.3. We then claim that if $\mathrm{kl}(\widehat{\mu}_{1,s}, \theta_\delta) \gtrsim f_s(1)/s$, then for $s \leq M = \lceil \log(1/\delta)/(\log\log(1/\delta))^4 \rceil$, we also have $\mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) \gtrsim f_s(1)/(2s)$. To see this, set $y = f_s(1)/s$, $q = \mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1)$, $a = \varepsilon^2/V_{\min}$ and $b = \varepsilon\sqrt{2V}/V_{\min}$. Equation (36) gives $y \leq q + a + b\sqrt{q}$. Uniformly over $s \leq M$, $y \to \infty$. If $q < y/2$, then $q + a + b\sqrt{q} < y/2 + a + b\sqrt{y/2} < y$ for all sufficiently small δ, a contradiction. Thus $q \geq y/2$.

Moreover, since $\theta_\delta \leq \mu_1 + \varepsilon$ and ε is fixed, the quantity

$$\sup_{x \in [\mu_1, \theta_\delta]} \mathrm{kl}(x, \mu_1)$$

is finite. On the other hand, for $s \leq M$, $f_s(1)/s \geq \Theta((\log\log(1/\delta))^4) \to \infty$. Therefore, for all sufficiently small δ, the event

$$\{\widehat{\mu}_{1,s} \leq \theta_\delta,\, \mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) > f_s(1)/(2s)\}$$

implies $\widehat{\mu}_{1,s} \leq \mu_1$. Continuing from (35), we have

$$\begin{aligned}
\text{r.h.s. of (35)} &\lesssim \sum_{s=1}^{M} \mathbb{P}\Big(\widehat{\mu}_{1,s} \leq \mu_1,\, \mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) > \frac{f_s(1)}{2s}\Big) \\
&\leq \sum_{s=1}^{M} \exp\left(-s \cdot \frac{f_s(1)}{2s}\right) \\
&\lesssim \frac{1}{K\log^3(1/\delta)}. \qquad (37)
\end{aligned}$$

For the last inequality in (37), write $L = \log(1/\delta)$. Since $f_s(1) \geq \log(QL^2/\delta)$ and $M = \lceil L/(\log L)^4 \rceil$,

$$\sum_{s=1}^{M} e^{-f_s(1)/2} \leq M\sqrt{\frac{\delta}{QL^2}} = O\left(\frac{\sqrt{\delta}}{(\log L)^4}\right) = o\left(\frac{1}{KL^3}\right).$$

Let

$$E_1 = \left\{\exists\, t \leq T:\ N_1(t) \leq M,\, U_1(t) \leq \theta_\delta\right\}.$$

From (37),

$$\mathbb{P}(E_1) \lesssim \frac{1}{K\log^3(1/\delta)}. \tag{38}$$

Then, we have

$$\begin{aligned}
I_2 &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta, E_1^c\}\right] + T\cdot\mathbb{P}(E_1) \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta, E_1^c\}\right] + 1.
\end{aligned}$$

We note that, since $b_t = \arg\max_j U_j(t)$,

$$E_1^c \cap \{U_{b_t}(t) < \theta_\delta\} \subseteq \{U_1(t) < \theta_\delta\} \cap E_1^c \subseteq \{N_1(t) > M\}.$$

We have

$$\begin{aligned}
\mathbb{P}\left(\exists s \geq M:\ \widehat{\mu}_{1,s} \leq \mu_1 - \zeta_1\right) &\leq \exp\left(-M\cdot\zeta_1^2/(2V)\right) && \text{(Due to Lemma B.4)} \\
&\lesssim \frac{1}{K\log^2(1/\delta)}. && (39)
\end{aligned}$$

For the last inequality in (39), $\zeta_1^2 = 1/\log L$ and $M \geq L/(\log L)^4$, so

$$KL^2 e^{-M\zeta_1^2/(2V)} \leq \exp\left(\log K + 2\log L - \frac{L}{2V(\log L)^5}\right) \longrightarrow 0.$$

Let

$$E_2 = \left\{\exists s \geq M:\ \widehat{\mu}_{1,s} \leq \mu_1 - \zeta_1\right\}.$$

Then $\mathbb{P}(E_2) \lesssim 1/(K\log^2(1/\delta))$. We have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta, E_1^c\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta, N_1(t) > M\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta, N_1(t) > M, E_2^c\}\right] + T\cdot\mathbb{P}(E_2) \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, U_{b_t}(t) < \theta_\delta, N_1(t) > M, E_2^c\}\right] + \frac{QV}{\varepsilon^2}.
\end{aligned}$$

On the event inside the last expectation, we have $\widehat{\mu}_1(t) > \mu_1 - \zeta_1$ and

$$U_{b_t}(t) - \varepsilon + \frac{7\zeta_1}{2} < \theta_\delta - \varepsilon + \frac{7\zeta_1}{2} = \mu_1 - \frac{3\zeta_1}{2} < \widehat{\mu}_1(t).$$

Thus $1 \in \mathcal{C}_t$. Since $A_{t+1} = i$ and G_t holds, the update rule of Algorithm 1 implies

$$\widehat{\mu}_i(t) \geq \widehat{\mu}_1(t) - \zeta_1 \geq \mu_1 - 2\zeta_1.$$

Indeed, if the greedy update is triggered, then $\mathrm{a}^*(t)$ maximizes the empirical mean over $\mathcal{C}_t$; otherwise the greedy update condition fails, so $\max_{j\in\mathcal{C}_t} \widehat{\mu}_j(t) < \widehat{\mu}_{\mathrm{a}^*(t)}(t) + \zeta_1$. Therefore,

$$\begin{aligned}
I_2 &\lesssim \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{1}\{\widehat{\mu}_{i,s} \geq \mu_1 - 2\zeta_1\}\right] + \frac{QV}{\varepsilon^2} + 1 \\
&\leq \sum_{s=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i,s} \geq \mu_1 - 2\zeta_1\right) + \frac{QV}{\varepsilon^2} + 1 \\
&\leq \sum_{s=1}^{T} \exp\left(-\frac{s(\Delta_i - 2\zeta_1)^2}{2V}\right) + \frac{QV}{\varepsilon^2} + 1 = O(1).
\end{aligned}$$

Bounding Term I_3.

$$\begin{aligned}
I_3 &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{B}_t\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{B}_t, E_1^c, E_2^c\}\right] + T \cdot \mathbb{P}(E_1) + T \cdot \mathbb{P}(E_2) \\
&\lesssim 1 + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{B}_t, U_{b_t}(t) \geq \theta_\delta, E_1^c, E_2^c\}\right]}_{I_{3,1}} \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_{t+1} = i, \mathsf{B}_t, U_{b_t}(t) < \theta_\delta, E_1^c, E_2^c\}\right]}_{I_{3,2}} + \frac{2QV}{\varepsilon^2}.
\end{aligned}$$

For $I_{3,1}$, define

$$m_i := \left\lceil \frac{\log(Q\log^2(1/\delta)/\delta)}{\mathrm{kl}(\mu_i + \zeta_1, \theta_\delta)} \right\rceil.$$

Then

$$\begin{aligned}
I_{3,1} &\leq \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{1}\{U_{i,s} \geq \theta_\delta\}\right] \\
&\leq \mathbb{E}\left[\sum_{s=1}^{m_i} \mathbb{1}\{U_{i,s} \geq \theta_\delta\}\right] + \mathbb{E}\left[\sum_{s=m_i}^{T} \mathbb{1}\{U_{i,s} \geq \theta_\delta\}\right] \\
&\lesssim \frac{\log(Q\log^2(1/\delta)/\delta)}{\mathrm{kl}(\mu_i + \zeta_1, \theta_\delta)} + 1 + \mathbb{E}\left[\sum_{s=m_i}^{T} \mathbb{1}\{\widehat{\mu}_{i,s} \geq \mu_i + \zeta_1\}\right] \\
&= \frac{\log(Q\log^2(1/\delta)/\delta)}{\mathrm{kl}(\mu_i + \zeta_1, \theta_\delta)} + O(1),
\end{aligned}$$

For the third inequality, put $A_\delta = \log(Q\log^2(1/\delta)/\delta)$. For sufficiently small δ, $m_i\varepsilon^2/V \geq 1$. Therefore, for every integer $s \geq m_i$, including $s = m_i$,

$$f_s(1) = A_\delta, \qquad \frac{f_s(1)}{s} = \frac{A_\delta}{s} \leq \frac{A_\delta}{m_i} \leq \mathrm{kl}(\mu_i + \zeta_1, \theta_\delta).$$

Thus, if $U_{i,s} \geq \theta_\delta$, then necessarily $\widehat{\mu}_{i,s} \geq \mu_i + \zeta_1$; the remaining summation is $O(1)$ by Lemma B.4.

For $I_{3,2}$, assume that $U_{b_t}(t) < \theta_\delta$, E_1^c, and E_2^c are true. First, if $N_1(t) \le M$, then since E_1^c is true, we must have $U_1(t) > \theta_\delta$. Hence

$$U_{b_t}(t) \ge U_1(t) > \theta_\delta,$$

which contradicts $U_{b_t}(t) < \theta_\delta$. Second, if $N_1(t) > M$, then since E_2^c is true,

$$\widehat{\mu}_1(t) > \mu_1 - \zeta_1 > \theta_\delta - \varepsilon + \frac{7\zeta_1}{2} > U_{b_t}(t) - \varepsilon + \frac{7\zeta_1}{2}.$$

Thus $1 \in \mathcal{C}_t$, so the UCB branch B_t cannot occur. Therefore, $I_{3,2} = 0$. Thus, we have

$$I_3 \lesssim \frac{\log(Q\log^2(1/\delta)/\delta)}{\mathrm{kl}(\mu_i + \zeta_1, \theta_\delta)} + O(1).$$

Substituting I_1, I_2, and I_3 into (33), we have

$$\mathbb{E}_{\boldsymbol{\mu}}[N_i[\tau_\delta]] \le \frac{\log(Q\log^2(1/\delta)/\delta)}{\mathrm{kl}(\mu_i + \zeta_1, \theta_\delta)} + O(1).$$

Since $\theta_\delta = \mu_1 + \varepsilon - 5\zeta_1$ and $\zeta_1 \to 0$, we obtain

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\tau_\delta)]}{\log(1/\delta)} \le \sum_{i>1} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu_1 + \varepsilon)}.$$

D.2 Proof of the Asymptotic Sample Complexity ((13) in Theorem 5.1)

By the construction of `KLUCB+Greedy`, the algorithm stops no later than the deterministic horizon

$$T = 3K + \left\lceil 32KV\zeta^{-2}\log\left(\frac{QK\log^2(1/\bar{\delta})}{\bar{\delta}}\right)\right\rceil.$$

Therefore,

$$\mathbb{E}_\mu[\tau_\delta] \le T.$$

Recall that

$$\zeta = \min\left\{\frac{1}{\sqrt{\log\log(1/\bar{\delta})}}, \frac{\epsilon}{4}\right\}.$$

Since $\bar{\delta} = \delta$ for all sufficiently small δ, we have

$$\zeta^{-2} = \log\log(1/\delta).$$

Hence,

$$\mathbb{E}_\mu[\tau_\delta] \le 3K + 1 + 32KV\log\log(1/\delta)\log\left(\frac{QK\log^2(1/\delta)}{\delta}\right) = O\big(\log(1/\delta)\log\log(1/\delta)\big).$$

Therefore,

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

D.3 Proof of Correctness and ε-PAC* property

The proof of this part requires the following two lemmas. The first lemma states that, with high probability, the upper confidence bound of the optimal arm never drops below μ_1 by more than ζ.

Lemma D.1. *Let T be as defined in Theorem 5.1. Then,*

$$\mathbb{P}\big(\exists\, t \le T : U_1(t) \le \mu_1 - \zeta\big) \le \frac{\delta}{4}.$$

The second lemma states that, with high probability, no lower confidence bound exceeds its true mean by more than ζ.

Lemma D.2. *Let T be as defined in Theorem 5.1. Then,*

$$\mathbb{P}\Big(\exists\, i \in [K],\ \exists\, t \le T : L_i(t) \ge \mu_i + \zeta\Big) \le \frac{\delta}{4}.$$

Now, we are ready to prove the correctness of Algorithm 1. We first show that, whenever the stopping rule is triggered, the returned arm is an ε-best arm with probability at least $1 - \delta/2$. We then show that, with probability at least $1 - \delta/2$, the algorithm stops before the deterministic horizon T.

We first prove the first claim. By Lemmas D.1 and D.2, with probability at least $1-\delta/2$, the following two inequalities hold simultaneously for all $t \le T$:

$$U_1(t) \ge \mu_1 - \zeta, \qquad L_i(t) \le \mu_i + \zeta, \quad \forall i \in [K].$$

On this event, suppose that the stopping rule is satisfied at time t. If $\mathrm{a}^*(t) = 1$, then the returned arm is clearly ε-best. Otherwise, $\mathrm{a}^*(t) \ne 1$, and the stopping rule gives

$$\begin{aligned} L_{\mathrm{a}^*(t)}(t) &\ge \max_{i \in [K]\setminus\{\mathrm{a}^*(t)\}} U_i(t) - \varepsilon + 2\zeta \\ &\ge U_1(t) - \varepsilon + 2\zeta \\ &\ge \mu_1 - \varepsilon + \zeta. \end{aligned}$$

Together with $L_{\mathrm{a}^*(t)}(t) \le \mu_{\mathrm{a}^*(t)} + \zeta$, this implies

$$\mu_{\mathrm{a}^*(t)} \ge \mu_1 - \varepsilon.$$

Hence the returned arm is ε-best.

We now prove the second claim. Let

$$\mathcal{E} = \big\{\forall t \in [T],\, U_1(t) \ge \mu_1 - \zeta\big\}.$$

Let B_t denote the event that Algorithm 1 uses the UCB branch at time t, that is, $\mathcal{C}_t = \emptyset$ and $A_{t+1} = b_t$. Let G_t denote the event that the algorithm uses the candidate/greedy branch at time t, that is, $\mathcal{C}_t \ne \emptyset$ and $A_{t+1} = a^*(t)$. Then

$$\begin{aligned} &\mathbb{P}\Big(\sum_{i\in[K]} N_i[\tau_\delta] \ge T\Big) \\ &\le \mathbb{P}(\mathcal{E}^c) + \mathbb{P}\Big(\sum_{i\in[K]}\sum_{t=K}^{T-1} \mathbb{1}\{A_{t+1} = i, \mathcal{E}\} \ge T - K\Big) \\ &\le \frac{\delta}{4} + \mathbb{P}\Big(\sum_{i\in[K]}\sum_{t=K}^{T-1} \mathbb{1}\{A_{t+1} = i, \mathcal{E}\} \ge T - K\Big) && \text{(by Lemma D.1)} \\ &\le \frac{\delta}{4} + \underbrace{\mathbb{P}\Big(\sum_{i\in[K]}\sum_{t=K}^{T-1} \mathbb{1}\{A_{t+1} = i, \mathsf{B}_t, \mathcal{E}\} \ge \frac{T-K}{2}\Big)}_{\mathrm{W}_1} \\ &\quad + \underbrace{\mathbb{P}\Big(\sum_{i\in[K]}\sum_{t=K}^{T-1} \mathbb{1}\{A_{t+1} = i, \mathsf{G}_t, \mathcal{E}\} \ge \frac{T-K}{2}\Big)}_{\mathrm{W}_2}. \end{aligned} \tag{40}$$

Let

$$s_0 = \left\lceil \frac{8V \log(QK \log^2(1/\bar\delta)/\bar\delta)}{\zeta^2} \right\rceil.$$

By Lemma B.4 and a union bound over arms,

$$\mathbb{P}(\mathcal{H}^c) \le \frac{\delta}{32}, \qquad \mathcal{H} := \left\{\forall i \in [K],\ \forall s \ge s_0,\ |\widehat{\mu}_{i,s} - \mu_i| \le \frac{\zeta}{2}\right\}. \tag{41}$$

In the following, we work on the event $\mathcal{E} \cap \mathcal{H}$. The inequality $\varepsilon \geq 4\zeta$ holds for every $\delta \in (0,1)$. Also, $s \geq s_0$ implies $s\varepsilon^2/V \geq 1$ and

$$\frac{f_s(K)}{s} \leq \frac{\zeta^2}{8V}.$$

By the definition of T, $T - K > 2Ks_0$.

Bounding Term W_1**.** The UCB branch requires $\mathcal{C}_t = \emptyset$ and chooses $b_t \in \arg\max_i U_i(t)$. If $N_{b_t}(t) \geq s_0$, Lemma B.3 and the definition of s_0 give

$$U_{b_t}(t) \leq \widehat{\mu}_{b_t}(t) + \zeta/2.$$

If the comparison mean lies outside I, the clipped index gives the same bound. Since $\varepsilon \geq 4\zeta$,

$$\widehat{\mu}_{b_t}(t) \geq U_{b_t}(t) - \zeta/2 \geq U_{b_t}(t) - \varepsilon + 7\zeta/2.$$

Thus $b_t \in \mathcal{C}_t$, a contradiction. Hence every UCB-branch pull selects an arm with fewer than s_0 observations. There are fewer than $Ks_0 < (T-K)/2$ such pulls, so $\mathrm{W}_1 = 0$.

Bounding Term W_2**.** We next show that, on $\mathcal{E} \cap \mathcal{H}$, the candidate/greedy branch also cannot be used at least $(T-K)/2$ times without stopping. Suppose that at time t, $A_{t+1} = a^*(t) = i$, G_t occurs, and $N_i(t) = s \geq s_0$. By the definition of s_0,

$$\mathrm{kl}(\widehat{\mu}_{i,s}, \widehat{\mu}_{i,s} - \zeta/2) \geq \frac{\zeta^2}{8V} \geq \frac{f_s(K)}{s}.$$

Therefore,

$$L_i(t) \geq \widehat{\mu}_{i,s} - \frac{\zeta}{2}.$$

Since G_t occurs and $a^*(t) = i$, the update rule of Algorithm 1 implies

$$\widehat{\mu}_i(t) \geq U_{b_t}(t) - \varepsilon + \frac{7}{2}\zeta - \zeta.$$

Thus

$$L_i(t) \geq \widehat{\mu}_i(t) - \frac{\zeta}{2} \geq U_{b_t}(t) - \varepsilon + 2\zeta \geq \max_{j \in [K]\setminus\{i\}} U_j(t) - \varepsilon + 2\zeta.$$

Hence the stopping rule is triggered at time t. Therefore, on $\mathcal{E} \cap \mathcal{H}$, before the algorithm stops, each arm can be selected by the candidate/greedy branch fewer than s_0 times. Since $Ks_0 < (T-K)/2$, we get

$$\mathrm{W}_2 \leq \mathbb{P}(\mathcal{H}^c) \leq \frac{\delta}{32}.$$

Substituting the bounds for W_1 and W_2 into (40), we obtain

$$\mathbb{P}\Big(\sum_{i \in [K]} N_i[\tau_\delta] \geq T \Big) \leq \frac{\delta}{2}. \tag{42}$$

Equivalently, with probability at least $1 - \delta/2$, the algorithm stops before the deterministic horizon T. This completes the proof.

Proof of the ε-PAC* property. It remains to verify the additional requirement in Definition 4.4. Recall that arm 1 is the unique best arm and define

$$\Delta_{\min} := \mu_1 - \max_{i \neq 1} \mu_i > 0.$$

Since $\zeta \to 0$ as $\delta \downarrow 0$, for all sufficiently small δ, $\zeta \leq \frac{\Delta_{\min}}{4}$. Also, by the definition of ζ, we have $\zeta \leq \varepsilon/4$, and hence $\varepsilon \geq 4\zeta$. Let

$$M_\delta := \left\lceil \frac{\log(1/\delta)}{\big(\log\log(1/\delta)\big)^4} \right\rceil.$$

This choice satisfies $M_\delta = o(T)$ and $M_\delta \zeta^2 \to \infty$ as $\delta \downarrow 0$. Define the following events:

$$\mathcal{E}_U := \big\{ \forall t \leq T,\ U_1(t) \geq \mu_1 - \zeta \big\},$$

$$\mathcal{E}_L := \big\{\forall i \in [K],\ \forall t \le T,\ L_i(t) \le \mu_i + \zeta\big\},$$
$$\mathcal{E}_{\text{stop}} := \{\tau_\delta < T\},$$

and

$$\mathcal{E}_\mu := \big\{\forall i \in [K],\ \forall s \ge M_\delta,\ |\widehat{\mu}_{i,s} - \mu_i| \le \zeta\big\}.$$

Here $L_i(t)$ denotes the same lower-confidence index as in Algorithm 1, computed for arm i.

Fix $H \in I$ with $H > \mu_1 + \varepsilon$. For each arm $i \in [K]$, define

$$\mathcal{E}_i^u := \big\{\exists t \le T : N_i(t) \le M_\delta,\ U_i(t) \le H\big\},$$

and, for each arm $i \ne 1$, define

$$\mathcal{E}_i^\ell := \big\{\exists t \le T : N_i(t) \le M_\delta,\ L_i(t) \ge \mu_i - \varepsilon\big\}.$$

By Lemmas D.1 and D.2, and (42),

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_U^c) \le \delta/4, \qquad \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_L^c) \le \delta/4, \qquad \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_{\text{stop}}^c) \le \delta/2.$$

We next control the small-sample events. If $\mathcal{E}_i^u$ occurs, then for some $s = N_i(t) \le M_\delta$, we have

$$\widehat{\mu}_{i,s} \le H, \qquad s\mathrm{kl}\big(\widehat{\mu}_{i,s}, H\big) \ge f_s(1).$$

Similarly, if $\mathcal{E}_i^\ell$ occurs for $i \ne 1$, then for some $s = N_i(t) \le M_\delta$ we have

$$\widehat{\mu}_{i,s} \ge \mu_i - \varepsilon, \qquad s\mathrm{kl}\big(\widehat{\mu}_{i,s}, \mu_i - \varepsilon\big) \ge f_s(K).$$

For $\mathcal{E}_i^u$, apply the argument leading to (37) on the compact interval $[\mu_i, H]$, with $H - \mu_i$ in place of ε. For $\mathcal{E}_i^\ell$, use the same argument and the upper-tail maximal inequality on $[\mu_i - \varepsilon, \mu_i]$. Since $M_\delta = o(\log(1/\delta))$, both arguments give

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i^u) \lesssim \frac{1}{2K\log^3(1/\delta)}, \qquad \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i^\ell) \lesssim \frac{1}{2K\log^3(1/\delta)} \quad (i \ne 1).$$

For the empirical-mean event, Lemma B.4 and the union bound over arms imply

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_\mu^c) \le 2K\exp\left(-\frac{M_\delta \zeta^2}{2V}\right) = o(1),$$

because $M_\delta \zeta^2 \to \infty$. Therefore, for

$$\mathcal{G}_\delta := \mathcal{E}_U \cap \mathcal{E}_L \cap \mathcal{E}_{\text{stop}} \cap \mathcal{E}_\mu \cap \bigcap_{i=1}^K (\mathcal{E}_i^u)^c \cap \bigcap_{i \ne 1} (\mathcal{E}_i^\ell)^c,$$

we have

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{G}_\delta) \longrightarrow 1 \qquad \text{as } \delta \downarrow 0.$$

We now prove that on $\mathcal{G}_\delta$ the algorithm returns arm 1. Fix a sample path in $\mathcal{G}_\delta$ and let $t := \tau_\delta$. Since $\mathcal{E}_{\text{stop}}$ holds, $t < T$, so the output is generated by the stopping rule. Hence $i_{\text{out},\delta} = \mathrm{a}^*(t)$. Suppose, for contradiction, that $\mathrm{a}^*(t) = i \ne 1$. The stopping rule gives

$$L_i(t) \ge \max_{j \in [K]\setminus\{i\}} U_j(t) - \varepsilon + 2\zeta \ge U_1(t) - \varepsilon + 2\zeta. \tag{43}$$

We first rule out the case $N_i(t) \le M_\delta$. Since $(\mathcal{E}_i^\ell)^c$ holds, $L_i(t) < \mu_i - \varepsilon$. Because $i \ne 1$, $\mu_i \le \mu_1 - \Delta_{\min}$. Thus

$$L_i(t) < \mu_1 - \Delta_{\min} - \varepsilon \le \mu_1 - \varepsilon - 4\zeta.$$

On the other hand, by (43) and $\mathcal{E}_U$,

$$L_i(t) \ge U_1(t) - \varepsilon + 2\zeta \ge \mu_1 - \zeta - \varepsilon + 2\zeta = \mu_1 - \varepsilon + \zeta,$$

which is impossible. Therefore,

$$N_i(t) > M_\delta. \tag{44}$$

Next consider the case $N_1(t) \le M_\delta$. Since $(\mathcal{E}_1^u)^c$ holds, $U_1(t) > H > \mu_1 + \varepsilon$. Then the stopping rule implies

$$L_i(t) \ge U_1(t) - \varepsilon + 2\zeta > \mu_1 + 2\zeta.$$

But $\mathcal{E}_L$ gives

$$L_i(t) \le \mu_i + \zeta \le \mu_1 - \Delta_{\min} + \zeta \le \mu_1 - 3\zeta,$$

again impossible. Hence,

$$N_1(t) > M_\delta. \tag{45}$$

It remains to handle the case in which both (44) and (45) hold. Since $\mathcal{E}_\mu$ holds,

$$\widehat{\mu}_1(t) \ge \mu_1 - \zeta$$

and

$$\widehat{\mu}_i(t) \le \mu_i + \zeta \le \mu_1 - \Delta_{\min} + \zeta \le \mu_1 - 3\zeta.$$

Consequently,

$$\widehat{\mu}_i(t) \le \widehat{\mu}_1(t) - 2\zeta. \tag{46}$$

We now show that arm 1 must belong to the candidate set $\mathcal{C}_t$. Recall that $b_t = \arg\max_{j\in[K]} U_j(t)$. and

$$\mathcal{C}_t = \left\{ j \in [K] : \widehat{\mu}_j(t) \ge U_{b_t}(t) - \varepsilon + \frac{7}{2}\zeta \right\}.$$

We distinguish two cases.

First, suppose $b_t = i$. Consider the last pull of arm 1. Write $L = \log(1/\delta)$ and

$$q_s := \frac{f_s(1)}{s}.$$

For all sufficiently small δ, $M_\delta \varepsilon^2 / V \ge 1$, so $f_s(1) = A_\delta := \log(Q \log^2(1/\delta)/\delta)$ for $s \ge M_\delta$. Consequently, q_s is nonincreasing and

$$0 \le q_s - q_{s+1} = \frac{A_\delta}{s(s+1)} \le \frac{A_\delta}{M_\delta^2} = o(1), \qquad s \ge M_\delta.$$

We use the following continuity property. Let

$$\mathsf{u}(x, q) := \sup\{y \in I : y \ge x,\ \mathrm{kl}(x, y) \le q\}.$$

For a fixed interior mean μ, if $x_\delta, y_\delta \to \mu$, $q'_\delta \le q_\delta + o(1)$, and $\mathsf{u}(x_\delta, q_\delta) \le B$ for a fixed $B \in I$ with $B > \mu$, then

$$\mathsf{u}(y_\delta, q'_\delta) \le \mathsf{u}(x_\delta, q_\delta) + o(1). \tag{47}$$

Indeed, $q_\delta \le \mathrm{kl}(x_\delta, B)$ is bounded. Choose an interior $B' > B$. Continuity of kl and its strict increase in its second argument above x keep the perturbed upper endpoints below B' and give uniform continuity of the inverse on this compact set, including $q = 0$. The bound is uniform on these compact sets. It also holds for $q'_\delta < q_\delta$ because $\mathsf{u}(x, q)$ is nondecreasing in q. Thus the $o(1)$ terms below are uniform on $\mathcal{G}_\delta$.

Let $u \le t$ be the last round in which arm 1 was actually pulled, and use superscripts $-$ and $+$ for the statistics immediately before and after that pull. Its pre-pull count is

$$N_1^- = N_1(t) - 1 \ge M_\delta.$$

There are two possibilities.

If this pull used the candidate/greedy branch, its selected incumbent was arm 1. The update rule gives

$$\widehat{\mu}_1^- \ge \max_j U_j^- - \varepsilon + \frac{5\zeta}{2}.$$

Since $\mathcal{E}_\mu$ holds,

$$\max_j U_j^- \le \mu_1 + \varepsilon - \frac{3\zeta}{2} < H.$$

The complements of the small-sample events $\mathcal{E}_j^u$ therefore imply $N_j^- > M_\delta$ for every arm j. From this round onward, all empirical means satisfy the bounds in $\mathcal{E}_\mu$, so arm 1 is the unique empirical best arm, separated from every other arm by at least 2ζ. The update rule then keeps arm 1: when the candidate set is nonempty it contains arm 1, and no other arm has a larger empirical mean; when it is empty, the incumbent is unchanged. Hence the final incumbent is arm 1, a contradiction.

It remains to consider a last pull of arm 1 on the UCB branch. Immediately before that pull, $U_i^- \le U_1^-$. Since arm 1 is never sampled again, $U_1^+ = U_1(t)$. The stopping rule and $\mathcal{E}_\mu$ give

$$U_1^+ \le L_i(t) + \varepsilon - 2\zeta \le \widehat{\mu}_i(t) + \varepsilon - 2\zeta \le \mu_i + \varepsilon - \zeta < H.$$

Apply (47) to the two consecutive empirical means of arm 1, which both lie within ζ of μ_1, and use $q_{N_1^-} - q_{N_1^-+1} = o(1)$. We obtain

$$U_1^- \le U_1^+ + o(1) \le \mu_i + \varepsilon + o(1) < H.$$

Thus $U_i^- \le U_1^- < H$, and $(\mathcal{E}_i^u)^c$ implies $N_i^- > M_\delta$. From that time to stopping, both empirical means of arm i lie within ζ of μ_i, while its confidence threshold q_s can only decrease. A second application of (47) yields

$$U_i(t) \le U_i^- + o(1) \le U_1(t) + o(1).$$

Therefore,

$$\begin{aligned} U_{b_t}(t) - \varepsilon + \frac{7\zeta}{2} &= U_i(t) - \varepsilon + \frac{7\zeta}{2} \\ &\le L_i(t) + \frac{3\zeta}{2} + o(1) \\ &\le \mu_i + \frac{5\zeta}{2} + o(1) < \mu_1 - \zeta \le \widehat{\mu}_1(t) \end{aligned}$$

for all sufficiently small δ, because the fixed gap $\mu_1 - \mu_i$ is positive. Hence $1 \in \mathcal{C}_t$ also in this case.

Second, suppose $b_t \neq i$. Since $b_t \in [K] \setminus \{i\}$, the stopping rule implies

$$L_i(t) \ge U_{b_t}(t) - \varepsilon + 2\zeta.$$

Using $L_i(t) \le \widehat{\mu}_i(t)$, we get

$$U_{b_t}(t) - \varepsilon + \frac{7}{2}\zeta \le L_i(t) + \frac{3}{2}\zeta \le \widehat{\mu}_i(t) + \frac{3}{2}\zeta \le \widehat{\mu}_1(t) - \frac{1}{2}\zeta < \widehat{\mu}_1(t),$$

where the penultimate inequality follows from (46). Hence $1 \in \mathcal{C}_t$ in this case as well.

Thus, in all cases,

$$1 \in \mathcal{C}_t. \tag{48}$$

Since $\mathrm{a}^*(t) = i \neq 1$ and $1 \in \mathcal{C}_t$, the update rule of Algorithm 1 yields a contradiction. Indeed, if the greedy step is triggered at time t, then $\mathrm{a}^*(t) = \arg\max_{j \in \mathcal{C}_t} \widehat{\mu}_j(t)$, and hence, by (48), $\widehat{\mu}_i(t) \ge \widehat{\mu}_1(t)$, which contradicts (46). If the greedy branch is not triggered, then the algorithm keeps the previous candidate, so $\mathrm{a}^*(t) = \mathrm{a}^*(t-1) = i$, and the failure of the greedy condition gives $\max_{j \in \mathcal{C}_t} \widehat{\mu}_j(t) < \widehat{\mu}_i(t) + \zeta$. Again using $1 \in \mathcal{C}_t$, we obtain $\widehat{\mu}_1(t) < \widehat{\mu}_i(t) + \zeta$, or equivalently, $\widehat{\mu}_i(t) > \widehat{\mu}_1(t) - \zeta$, which also contradicts (46). Therefore our assumption $\mathrm{a}^*(t) = i \neq 1$ is impossible.

We have shown that, on $\mathcal{G}_\delta$,

$$i_{\mathrm{out},\delta} = \mathrm{a}^*(\tau_\delta) = 1.$$

Therefore,

$$\mathbb{P}_{\boldsymbol{\mu}}(i_{\mathrm{out},\delta} = 1) \ge \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{G}_\delta) \longrightarrow 1 \qquad \text{as } \delta \downarrow 0.$$

This proves that the family of policies is ε-PAC*.

Proof of Supporting Lemmas of Section D.3.

Proof of Lemma D.1. Use the algorithmic constants $Q = 128e^4$, $L = \log(1/\bar{\delta})$ and $L \ge L_0 = \max\{32, 2\log K, \varepsilon^4\}$. Put

$$h = 32V/\zeta^2, \qquad B_x = \log(e^4 x L/\bar{\delta}), \qquad A_x = \log(Qx L^2/\bar{\delta}), \quad 1 \le x \le K.$$

The choice of L_0 gives

$$h\varepsilon^2/V = 32\max\{\varepsilon^2\log L, 16\} \le 128L, \qquad B_x + 3 \le 2L, \qquad A_x \le 2L.$$

Indeed, $\varepsilon^2 \le \sqrt{L}$, $\log L \le \sqrt{L}$, $\log K \le L/2$, $\log L + 7 \le L/2$, and $2\log L + \log Q \le L/2$ for $L \ge 32$.

Split the event at h:

$$\mathbb{P}(\exists t \le T : U_1(t) \le \mu_1 - \zeta) \le \underbrace{\mathbb{P}(\exists 1 \le s \le h : U_{1,s} \le \mu_1 - \zeta)}_{G_1} + \underbrace{\mathbb{P}(\exists h < s \le T : U_{1,s} \le \mu_1 - \zeta)}_{G_2}.$$

All sample indices s are integers; empty intervals contribute zero.

Bounding G_1. For any $\beta > 1$ and $s \in (h/\beta^{\ell+1}, h/\beta^{\ell}]$,

$$\min\{1, s\varepsilon^2/V\} \le 128L/\beta^{\ell}, \qquad f_s(x) \ge B_x + \ell\log\beta.$$

The lower-tail maximal inequality, applied on these geometric intervals, therefore gives

$$\begin{aligned}\mathbb{P}(\exists s \le h : \widehat{\mu}_{1,s} \le \mu_1,\ s\mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) \ge f_s(x)) \\ \le \sum_{\ell=0}^{\infty} \exp\left(-\frac{B_x + \ell\log\beta}{\beta}\right) = \frac{e^{-B_x/\beta}}{1 - e^{-\log\beta/\beta}}.\end{aligned}$$

Choose $\beta = B_x/(B_x - 1)$. Since $1/(1 - e^{-u}) \le 1 + 1/u$ for $u > 0$ and $\log\beta \ge (\beta - 1)/\beta$,

$$\frac{1}{1 - e^{-\log\beta/\beta}} \le 1 + \frac{\beta^2}{\beta - 1} \le B_x + 3.$$

Consequently,

$$\frac{e^{-B_x/\beta}}{1 - e^{-\log\beta/\beta}} \le \frac{(B_x + 3)\bar{\delta}}{e^3 xL} \le \frac{2\bar{\delta}}{e^3 x} < \frac{\bar{\delta}}{8x}.$$

Taking $x = 1$ gives $G_1 \le \bar{\delta}/8$.

Bounding G_2. For $y \le \mu_1 - \zeta$, Proposition B.1 and Lemma B.3 yield

$$\begin{aligned}\mathrm{kl}(y, \mu_1) - \mathrm{kl}(y, \mu_1 - \zeta) &= \int_{\mu_1 - \zeta}^{\mu_1} \frac{u - y}{V(u)}\,\mathrm{d}u \\ &\ge \mathrm{kl}(\mu_1 - \zeta, \mu_1) \ge \frac{\zeta^2}{2V}.\end{aligned}$$

Thus $U_{1,s} \le \mu_1 - \zeta$ implies $\widehat{\mu}_{1,s} \le \mu_1 - \zeta$ and

$$s\mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) \ge f_s(1) + s\zeta^2/(2V).$$

For a general confidence factor x, use $f_s(x) \ge A_x$ and peel the intervals $(h\alpha^{\ell}, h\alpha^{\ell+1}]$, with $\alpha = A_x/(A_x - 1)$. The maximal inequality gives

$$\begin{aligned}&\mathbb{P}\big(\exists h < s \le T : \widehat{\mu}_{1,s} \le \mu_1 - \zeta,\ s\mathrm{kl}(\widehat{\mu}_{1,s}, \mu_1) \ge f_s(x) + s\zeta^2/(2V)\big) \\ &\quad\le \sum_{\ell=0}^{\infty} e^{-A_x/\alpha - 16\alpha^{\ell}} \\ &\quad\le e^{-A_x/\alpha}\frac{e^{-16}}{1 - e^{-16(\alpha-1)}} \le A_x e^{-A_x/\alpha} = \frac{eA_x\bar{\delta}}{QxL^2} \le \frac{\bar{\delta}}{8x}.\end{aligned}$$

Here $\alpha^{\ell} \ge 1 + \ell(\alpha - 1)$, and the penultimate inequality uses $1/(1 - e^{-u}) \le 1 + 1/u$. Taking $x = 1$ proves $G_2 \le \bar{\delta}/8$. Therefore

$$\mathbb{P}(\exists t \le T : U_1(t) \le \mu_1 - \zeta) \le \bar{\delta}/4 \le \delta/4.$$

□

Proof of Lemma D.2. Fix an arm $i \in [K]$. We prove

$$\mathbb{P}\big(\exists t \le T:\ L_i(t) \ge \mu_i + \zeta\big) \le \frac{\delta}{4K},$$

and then take a union bound over $i \in [K]$.

Let $s = N_i(t)$. If $L_i(t) \ge \mu_i + \zeta$, then necessarily $\widehat{\mu}_{i,s} \ge \mu_i + \zeta$; otherwise the lower-confidence index, which is constrained to be no larger than $\widehat{\mu}_{i,s}$, cannot exceed $\mu_i + \zeta$. Moreover, by the definition of the lower-confidence index,

$$s\mathrm{kl}\big(\widehat{\mu}_{i,s}, \mu_i + \zeta\big) \ge f_s(K).$$

Split the sample indices at $h = 32V/\zeta^2$, as in the preceding proof. For $s \le h$, replacing the lower-tail maximal inequality by its upper-tail counterpart and using $x = K$ in the bound for G_1 gives probability at most $\bar{\delta}/(8K)$. For $s > h$ and $y \ge \mu_i + \zeta$,

$$\mathrm{kl}(y, \mu_i) - \mathrm{kl}(y, \mu_i + \zeta) = \int_{\mu_i}^{\mu_i + \zeta} \frac{y - u}{V(u)}\, \mathrm{d}u \ge \frac{\zeta^2}{2V}.$$

Hence the upper-tail version of the bound for G_2, again with $x = K$, is at most $\bar{\delta}/(8K)$. Therefore,

$$\mathbb{P}(\exists t \le T : L_i(t) \ge \mu_i + \zeta) \le \frac{\bar{\delta}}{4K} \le \frac{\delta}{4K}.$$

Taking a union bound over $i \in [K]$ proves the lemma. □

E On the Characteristic Time

As we have discussed the case $C_1(\boldsymbol{\mu})$ before, next we consider the case $j \neq 1$. Throughout this section, fix $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu}) \setminus \{1\}$. If $\Delta_j = \varepsilon$, then the arm–1 branch below has zero cost and hence $C_j(\boldsymbol{\mu}) = +\infty$; such an arm cannot decrease $C(\boldsymbol{\mu}) = \min_{i \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_i(\boldsymbol{\mu})$. Thus, when deriving the parametric formula, we focus on the nondegenerate case $0 < \Delta_j < \varepsilon$.

For any $\alpha \in [0, 1]$, define the *ε-shifted transport cost*

$$I_\alpha^{(\varepsilon)}(u, v) \ = \min_{z \in [\mu_1 - \varepsilon, u]} \Big\{ \alpha\, \mathrm{kl}(u, z) \ + (1 - \alpha)\, \mathrm{kl}\big(v, z + \varepsilon\big) \Big\}. \tag{49}$$

The closed interval is used because all characteristic constants are defined through infima over alternatives; replacing the strict alternative by its closure does not change the infimum.

For the branch in which arm 1 certifies that j is not ε-best, set

$$S_1 = \{\boldsymbol{\lambda} : \lambda_1 = \mu_1,\ \lambda_j \le \mu_1 - \varepsilon\}.$$

For every $a \notin \{j, 1\}$, set

$$S_a^+ = \{\boldsymbol{\lambda} : \lambda_1 = \mu_1,\ \lambda_j \ge \mu_1 - \varepsilon,\ \lambda_a \ge \lambda_j + \varepsilon\}.$$

The additional constraint $\lambda_j \ge \mu_1 - \varepsilon$ prevents the pair branch from double-counting alternatives already covered by S_1. For infimum calculations,

$$\overline{\mathrm{Alt}_j(\mu_1)} = S_1 \cup \bigcup_{a \notin \{j,1\}} S_a^+,$$

where the bar only records the harmless closure of the strict inequality in the notion "not ε-best."

Lemma E.1. *For every $\boldsymbol{w} \in \mathcal{W}$, every $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu}) \setminus \{1\}$, and every $a \notin \{j, 1\}$, we have*

$$\inf_{\boldsymbol{\lambda} \in S_a^+} \sum_{i=2}^K w_i \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i} = \Big(\frac{w_j}{\Delta_j} + \frac{w_a}{\Delta_a}\Big) I^{(\varepsilon)}_{\frac{w_j/\Delta_j}{w_j/\Delta_j + w_a/\Delta_a}}(\mu_j, \mu_a).$$

For S_1,

$$\inf_{\boldsymbol{\lambda} \in S_1} \sum_{i=2}^K w_i \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i} = \frac{w_j}{\Delta_j} \mathrm{kl}(\mu_j, \mu_1 - \varepsilon).$$

Consequently, if

$$\Psi_j(\boldsymbol{w}) = \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}_j(\mu_1)} \sum_{i=2}^{K} w_i \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i},$$

then

$$\Psi_j(\boldsymbol{w}) = \min\left\{ \frac{w_j}{\Delta_j}\mathrm{kl}(\mu_j, \mu_1 - \varepsilon),\ \min_{a\notin\{j,1\}} \left(\frac{w_j}{\Delta_j} + \frac{w_a}{\Delta_a}\right) I^{(\varepsilon)}_{\frac{w_j/\Delta_j}{w_j/\Delta_j + w_a/\Delta_a}}(\mu_j, \mu_a) \right\}. \tag{50}$$

Let

$$\boldsymbol{w}^*(\boldsymbol{\mu}; j) \in \arg\max_{\boldsymbol{w}\in\mathcal{W}}\ \Psi_j(\boldsymbol{w}). \tag{51}$$

The following lemma shows that all pair branches are equalized at an optimizer.

Lemma E.2. *Assume $K \geq 3$ and $0 < \Delta_j < \varepsilon$. At any maximizer $\boldsymbol{w}^*(\boldsymbol{\mu}; j)$ of* (51)*, the quantities*

$$\left(\frac{w_j}{\Delta_j} + \frac{w_a}{\Delta_a}\right) I^{(\varepsilon)}_{\frac{w_j/\Delta_j}{w_j/\Delta_j + w_a/\Delta_a}}(\mu_j, \mu_a) \tag{52}$$

are all equal for every $a \notin \{j, 1\}$.

For every $a \notin \{j, 1\}$, define

$$g^{(\varepsilon)}_{j,a}(x) = (1+x) I^{(\varepsilon)}_{1/(1+x)}(\mu_j, \mu_a) = \min_{z\in[\mu_1-\varepsilon, \mu_j]} \{\mathrm{kl}(\mu_j, z) + x\,\mathrm{kl}(\mu_a, z+\varepsilon)\}, \qquad x \geq 0.$$

The map $x \mapsto g^{(\varepsilon)}_{j,a}(x)$ is continuous and strictly increasing on $[0, \infty)$. Indeed, by the envelope theorem, if z_x is a minimizer, then

$$\frac{d}{dx} g^{(\varepsilon)}_{j,a}(x) = \mathrm{kl}(\mu_a, z_x + \varepsilon) > 0,$$

because $z_x + \varepsilon \geq \mu_1 > \mu_a$. Hence it has an inverse on its range; write

$$x_{j\to a}(y) = \left(g^{(\varepsilon)}_{j,a}\right)^{-1}(y).$$

For $y > 0$, set

$$F^{\varepsilon}_{\boldsymbol{\mu},j}(y) = \Delta_j + \sum_{a\notin\{j,1\}} \Delta_a\, x_{j\to a}(y),$$

with the convention that the sum is zero when $K = 2$.

Lemma E.3. *Define*

$$A_j = \mathrm{kl}(\mu_j, \mu_1 - \varepsilon), \qquad G(y) = \frac{y}{F^{\varepsilon}_{\boldsymbol{\mu},j}(y)}.$$

Then

$$\left(C_j(\boldsymbol{\mu})\right)^{-1} = \sup_{y>0} \min\left\{ G(y),\ \frac{A_j}{F^{\varepsilon}_{\boldsymbol{\mu},j}(y)} \right\}. \tag{53}$$

Moreover, whenever a maximizer $y^* > 0$ in (53) *exists, an optimal regret allocation is obtained by*

$$w^*_j(\boldsymbol{\mu}; j) = \frac{\Delta_j}{F^{\varepsilon}_{\boldsymbol{\mu},j}(y^*)}, \qquad w^*_a(\boldsymbol{\mu}; j) = \frac{\Delta_a\, x_{j\to a}(y^*)}{F^{\varepsilon}_{\boldsymbol{\mu},j}(y^*)} \quad (a \notin \{j, 1\}), \qquad w^*_1(\boldsymbol{\mu}; j) = 0. \tag{54}$$

When $K = 2$, the formula is understood with the empty sum convention and gives $w^*_2 = 1$.

Time complexity. To compute $w^*(\boldsymbol{\mu}; j)$, one maximizes the one-dimensional objective

$$H_j(y) = \min\left\{ \frac{y}{F^{\varepsilon}_{\boldsymbol{\mu},j}(y)},\ \frac{\mathrm{kl}(\mu_j, \mu_1 - \varepsilon)}{F^{\varepsilon}_{\boldsymbol{\mu},j}(y)} \right\}, \qquad y > 0.$$

For any candidate y, evaluating $F^{\varepsilon}_{\boldsymbol{\mu},j}(y)$ requires, for each competitor $a \notin \{j, 1\}$, solving the monotone equation $g^{(\varepsilon)}_{j,a}(x) = y$ for $x = x_{j\to a}(y)$. Each inner solve can be performed by bisection to accuracy η in $O(\log(1/\eta))$ iterations. An outer one-dimensional search over y to the same accuracy therefore gives a single allocation in $O(K[\log(1/\eta)]^2)$ basic iterations, and computing the allocations for all possible j costs $O(K^2[\log(1/\eta)]^2)$ iterations.

E.1 Some Specific Instances

Let $C(\boldsymbol{\mu}) = \min_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_j(\boldsymbol{\mu})$. In general, closed form expressions for $C(\boldsymbol{\mu})$ are not available. However, for several salient instances one can obtain explicit formulas. In what follows we specialize the general characterization from the previous section to two important cases— Gaussian rewards with known variance and the two-armed bandit ($K = 2$)—and derive closed form expressions for $C(\boldsymbol{\mu})$ in these settings.

Lemma E.4 (Gaussian: $j = 1$)**.** *Assume Gaussian rewards $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$ with known σ^2. One has*

$$C(\boldsymbol{\mu}) = C^*(\boldsymbol{\mu}) = \sum_{i \geq 2} \frac{2\sigma^2 \Delta_i}{(\Delta_i + \varepsilon)^2}.$$

Lemma E.5 (Two-Armed Bandit)**.** *Assume $K = 2$. One has*

$$C(\boldsymbol{\mu}) = \min\left\{ \frac{\Delta_2}{\mathrm{kl}(\mu_2, \mu_1 + \varepsilon)}, \frac{\Delta_2}{\mathrm{kl}(\max\{\mu_2, \mu_1 - \varepsilon\}, \mu_1 - \varepsilon)} \right\},$$

where the ratio with denominator zero is interpreted as $+\infty$.

Remark E.6. This behavior contrasts sharply with the optimal *sample complexity* for best-arm identification studied in Garivier and Kaufmann [13], where even for Gaussian rewards and the two-armed case the optimum is given only implicitly by an optimization problem and typically has no simple closed form.

One might wonder whether the minimum in $C(\boldsymbol{\mu}) = \min_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_j(\boldsymbol{\mu})$ is always attained at the best arm, i.e., whether $C(\boldsymbol{\mu}) = C^*(\boldsymbol{\mu})$ for any $\boldsymbol{\mu}$. The answer is *no*. In the following, we exhibit a concrete instance for which $C^*(\boldsymbol{\mu}) > C(\boldsymbol{\mu})$.

Example E.7. We consider the Bernoulli reward distributions. Let $\mu_1 = 0.020$, $\mu_2 = 0.018$, and $\varepsilon = 0.012$. Then

$$\mu_1 + \varepsilon = 0.032, \qquad \mu_1 - \varepsilon = 0.008, \qquad \Delta_2 = \mu_1 - \mu_2 = 0.002.$$

The characteristic regrets for $K = 2$ are

$$C_1(\boldsymbol{\mu}) = \frac{\Delta_2}{\mathrm{kl}(\mu_2, \mu_1 + \varepsilon)}, \qquad C_2(\boldsymbol{\mu}) = \frac{\Delta_2}{\mathrm{kl}(\mu_2, \mu_1 - \varepsilon)}.$$

Plugging in the numbers,

$$\mathrm{kl}(0.018, 0.032) \approx 3.7442004902 \times 10^{-3}, \qquad \mathrm{kl}(0.018, 0.008) \approx 4.6473173422 \times 10^{-3},$$

hence

$$C_1(\boldsymbol{\mu}) = \frac{0.002}{0.0037442004902} \approx 0.5341594301,$$
$$C_2(\boldsymbol{\mu}) = \frac{0.002}{0.0046473173422} \approx 0.4303558059.$$

Therefore $\min\{C_1(\boldsymbol{\mu}), C_2(\boldsymbol{\mu})\} = C_2(\boldsymbol{\mu})$ and the minimizer is $j^* = 2 \neq 1$ for this instance.

E.2 Proofs of Supporting Lemmas

Proof of Proposition 4.3. By definition,

$$\frac{1}{C^*(\boldsymbol{\mu})} = \sup_{\boldsymbol{w} \in \mathcal{W}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}_1(\mu_1)} \sum_{i=2}^{K} w_i \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i},$$

where $w_1 = 0$ and $\sum_{i \geq 2} w_i = 1$. To make arm 1 not ε-best while keeping $\lambda_1 = \mu_1$, it is enough and optimal to choose one arm $i \geq 2$ and move only that arm to the boundary $\lambda_i = \mu_1 + \varepsilon$. All other coordinates remain equal to their original means because each KL term is nonnegative and uniquely minimized at the original mean. Therefore

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}_1(\mu_1)} \sum_{i=2}^{K} w_i \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i} = \min_{i \geq 2} w_i \frac{\mathrm{kl}(\mu_i, \mu_1 + \varepsilon)}{\Delta_i}.$$

Let $c_i = \mathrm{kl}(\mu_i, \mu_1 + \varepsilon)/\Delta_i$. The optimization becomes

$$\max_{\boldsymbol{w} \in \mathcal{W}} \min_{i \geq 2} w_i c_i.$$

Introducing a scalar t, this is the linear program

$$\begin{aligned} &\text{maximize } t \\ &\text{subject to } w_i c_i \geq t \quad (i \geq 2), \\ &\qquad \sum_{i \geq 2} w_i = 1, \quad w_i \geq 0. \end{aligned}$$

Since $w_i \geq t/c_i$, any feasible point satisfies $1 = \sum_{i \geq 2} w_i \geq t \sum_{i \geq 2} 1/c_i$, so $t \leq (\sum_{i \geq 2} 1/c_i)^{-1}$. Equality is achieved by

$$w_i^* = \frac{1/c_i}{\sum_{k \geq 2} 1/c_k},$$

for which all $w_i^* c_i$ are equal. Thus

$$\frac{1}{C_1(\boldsymbol{\mu})} = \left(\sum_{i \geq 2} \frac{\Delta_i}{\mathrm{kl}(\mu_i, \mu_1 + \varepsilon)} \right)^{-1},$$

which proves the stated formula for $C^*(\boldsymbol{\mu}) = C_1(\boldsymbol{\mu})$. □

Proof of Lemma E.1. Let

$$\Phi(\boldsymbol{\lambda}) = \sum_{i=2}^{K} w_i \frac{\mathrm{kl}(\mu_i, \lambda_i)}{\Delta_i}.$$

Because replacing strict inequalities by boundary inequalities does not change an infimum,

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}_j(\mu_1)} \Phi(\boldsymbol{\lambda}) = \min \Big\{ \inf_{\boldsymbol{\lambda} \in S_1} \Phi(\boldsymbol{\lambda}), \min_{a \notin \{j,1\}} \inf_{\boldsymbol{\lambda} \in S_a^+} \Phi(\boldsymbol{\lambda}) \Big\}.$$

For S_1, the constraint is $\lambda_j \leq \mu_1 - \varepsilon$. Since $\mu_j \geq \mu_1 - \varepsilon$ and $z \mapsto \mathrm{kl}(\mu_j, z)$ is decreasing for $z \leq \mu_j$, the minimum is attained at $\lambda_j = \mu_1 - \varepsilon$, while all other coordinates remain at their original means. This gives the displayed formula for the S_1 branch.

Now fix $a \notin \{j, 1\}$. For any coordinate $i \notin \{j, a\}$, the term $\mathrm{kl}(\mu_i, \lambda_i)$ is minimized by $\lambda_i = \mu_i$, so only coordinates j and a need to move. Write $z = \lambda_j$. In S_a^+, we have $z \geq \mu_1 - \varepsilon$. If $z > \mu_j$, decreasing z to μ_j decreases $\mathrm{kl}(\mu_j, z)$ and also decreases the boundary value $z + \varepsilon$, which is above $\mu_1 \geq \mu_a$; hence it cannot increase the arm-a KL term. Thus the optimizer has $z \in [\mu_1 - \varepsilon, \mu_j]$. For such z, $z + \varepsilon \geq \mu_1 > \mu_a$, so $\lambda_a \mapsto \mathrm{kl}(\mu_a, \lambda_a)$ is increasing on the feasible region $[z + \varepsilon, \infty)$; hence the boundary $\lambda_a = z + \varepsilon$ is optimal. Therefore

$$\inf_{\boldsymbol{\lambda} \in S_a^+} \Phi(\boldsymbol{\lambda}) = \min_{z \in [\mu_1 - \varepsilon, \mu_j]} \left\{ \frac{w_j}{\Delta_j} \mathrm{kl}(\mu_j, z) + \frac{w_a}{\Delta_a} \mathrm{kl}(\mu_a, z + \varepsilon) \right\}.$$

Factoring out $w_j/\Delta_j + w_a/\Delta_a$ yields exactly the shifted transport cost in the lemma. Combining the S_1 and pair branches gives (50). □

Proof of Lemma E.2. Set

$$s_j = \frac{w_j}{\Delta_j}, \qquad s_a = \frac{w_a}{\Delta_a}, \qquad f_a(\boldsymbol{w}) = s_j g_{j,a}^{(\varepsilon)}\left(\frac{s_a}{s_j}\right), \qquad a \notin \{j, 1\}.$$

The arm–1 branch is $h_1(\boldsymbol{w}) = s_j \mathrm{kl}(\mu_j, \mu_1 - \varepsilon)$. By (50), the objective is

$$\Psi_j(\boldsymbol{w}) = \min\{h_1(\boldsymbol{w}), \min_{a \notin \{j,1\}} f_a(\boldsymbol{w})\}.$$

Let $z_a(x)$ be a minimizer in the definition of $g_{j,a}^{(\varepsilon)}(x)$. The envelope theorem gives

$$\big(g_{j,a}^{(\varepsilon)}\big)'(x) = \mathrm{kl}(\mu_a, z_a(x) + \varepsilon) > 0,$$

and

$$g_{j,a}^{(\varepsilon)}(x) - x\big(g_{j,a}^{(\varepsilon)}\big)'(x) = \mathrm{kl}(\mu_j, z_a(x)) \geq 0.$$

Hence f_a is nondecreasing in s_j and strictly increasing in s_a whenever $f_a > 0$.

Let $\boldsymbol{w}^*$ be a maximizer and define $m = \min_{a \notin \{j,1\}} f_a(\boldsymbol{w}^*)$. Suppose that, for some $b \notin \{j, 1\}$, $f_b(\boldsymbol{w}^*) > m$. For sufficiently small $\eta > 0$, move mass from w_b to w_j:

$$w_j^\eta = w_j^* + \eta, \qquad w_b^\eta = w_b^* - \eta, \qquad w_a^\eta = w_a^* \quad (a \notin \{j, b\}).$$

This keeps the allocation feasible. The branch h_1 increases because w_j increases. Every pair branch that attained the old minimum m also increases, because its s_j increases while its own s_a is unchanged. The branch f_b decreases continuously, and since it was strictly above m, choosing η small keeps it above the new minimum of the formerly active branches. Therefore $\Psi_j(\boldsymbol{w}^\eta) > \Psi_j(\boldsymbol{w}^*)$, contradicting optimality. Thus all pair branches must be equal at any maximizer. □

Proof of Lemma E.3. If $K = 2$, there are no pair branches and (50) gives

$$(C_2(\boldsymbol{\mu}))^{-1} = \frac{\mathrm{kl}(\mu_2, \mu_1 - \varepsilon)}{\Delta_2},$$

which is also obtained from (53) using the empty-sum convention. We therefore assume $K \geq 3$.

Fix any feasible $\boldsymbol{w}$ with $s_j = w_j/\Delta_j > 0$ and set $x_a = (w_a/\Delta_a)/s_j$ for $a \notin \{j, 1\}$. Since $\sum_{i \geq 2} w_i = 1$,

$$\frac{1}{s_j} = \Delta_j + \sum_{a \notin \{j,1\}} \Delta_a x_a.$$

Let $y = \min_{a \notin \{j,1\}} g_{j,a}^{(\varepsilon)}(x_a)$. Since each $g_{j,a}^{(\varepsilon)}$ is increasing, $x_{j \to a}(y) \leq x_a$ for every $a \notin \{j, 1\}$, and hence

$$F_{\boldsymbol{\mu},j}^{\varepsilon}(y) = \Delta_j + \sum_{a \notin \{j,1\}} \Delta_a x_{j \to a}(y) \leq \frac{1}{s_j}.$$

Using (50),

$$\Psi_j(\boldsymbol{w}) = s_j \min\{A_j, y\} \leq \frac{\min\{A_j, y\}}{F_{\boldsymbol{\mu},j}^{\varepsilon}(y)}.$$

Taking the supremum over $\boldsymbol{w}$ gives the upper bound in (53).

Conversely, fix $y > 0$ and define the allocation by (54). Then $\sum_i w_i = 1$, and the corresponding ratios satisfy

$$\frac{w_a/\Delta_a}{w_j/\Delta_j} = x_{j \to a}(y), \qquad a \notin \{j, 1\}.$$

Therefore every pair branch equals $y/F_{\boldsymbol{\mu},j}^{\varepsilon}(y)$, while the arm–1 branch equals $A_j/F_{\boldsymbol{\mu},j}^{\varepsilon}(y)$. Hence this allocation achieves

$$\min\left\{ \frac{y}{F_{\boldsymbol{\mu},j}^{\varepsilon}(y)}, \frac{A_j}{F_{\boldsymbol{\mu},j}^{\varepsilon}(y)} \right\}.$$

Taking the supremum over $y > 0$ proves (53). If the supremum is attained at y^*, the same construction gives the optimal allocation (54). □

Proof of Lemma E.4. For Gaussian rewards with known variance σ^2, the variance function is constant on $(\mu_1 - \varepsilon, \mu_1 + \varepsilon)$, so the constant in (6) is $\ell = 1$. Lemma 4.7 therefore gives $C^*(\boldsymbol{\mu}) \leq C_j(\boldsymbol{\mu})$ for every $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})$. Since $C^*(\boldsymbol{\mu}) = C_1(\boldsymbol{\mu})$ and $C(\boldsymbol{\mu}) = \min_{j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})} C_j(\boldsymbol{\mu})$, we get $C(\boldsymbol{\mu}) = C^*(\boldsymbol{\mu})$. Finally, Proposition 4.3 and $\mathrm{kl}(\mu_i, \mu_1 + \varepsilon) = (\Delta_i + \varepsilon)^2/(2\sigma^2)$ yield

$$C^*(\boldsymbol{\mu}) = \sum_{i \geq 2} \frac{2\sigma^2 \Delta_i}{(\Delta_i + \varepsilon)^2}.$$

□

Proof of Lemma E.5. For $j = 1$, Proposition 4.3 gives

$$C_1(\boldsymbol{\mu}) = \frac{\Delta_2}{\mathrm{kl}(\mu_2, \mu_1 + \varepsilon)}.$$

If arm 2 is not ε-best, i.e. $\mu_2 < \mu_1 - \varepsilon$, then $\mathcal{I}_\varepsilon(\boldsymbol{\mu}) = \{1\}$ and $C(\boldsymbol{\mu}) = C_1(\boldsymbol{\mu})$. This agrees with the stated formula because $\max\{\mu_2, \mu_1 - \varepsilon\} = \mu_1 - \varepsilon$ and the second denominator is zero, so the second ratio is $+\infty$.

If arm 2 is ε-best, then $\mu_2 \geq \mu_1 - \varepsilon$. For $j = 2$, the only alternative branch is the arm–1 branch; hence

$$\frac{1}{C_2(\boldsymbol{\mu})} = \frac{1}{\Delta_2}\mathrm{kl}(\mu_2, \mu_1 - \varepsilon), \qquad C_2(\boldsymbol{\mu}) = \frac{\Delta_2}{\mathrm{kl}(\mu_2, \mu_1 - \varepsilon)},$$

with the convention $C_2(\boldsymbol{\mu}) = +\infty$ when the denominator is zero. Taking $\min\{C_1(\boldsymbol{\mu}), C_2(\boldsymbol{\mu})\}$ proves the claimed expression. □

Proof of Lemma 4.7. Let

$$V_+ = \sup_{\mu \in (\mu_1 - \varepsilon, \mu_1 + \varepsilon)} V(\mu), \qquad V_- = \inf_{\mu \in (\mu_1 - \varepsilon, \mu_1 + \varepsilon)} V(\mu),$$

so that $V_+/V_- \leq \ell$. For any x, y in this interval,

$$\mathrm{kl}(x, y) = \int_x^y \frac{y - u}{V(u)}\, du$$

where the expression follows from the standard Bregman representation; in particular, the quadratic comparison

$$\frac{(x-y)^2}{2V_+} \leq \mathrm{kl}(x, y) \leq \frac{(x-y)^2}{2V_-}$$

holds. Hence, for every $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu}) \setminus \{1\}$,

$$\mathrm{kl}(\mu_j, \mu_1 - \varepsilon) \leq \ell\, \mathrm{kl}(\mu_j, \mu_1 + \varepsilon), \tag{55}$$

because $(\varepsilon - \Delta_j)^2 \leq (\varepsilon + \Delta_j)^2$.

Fix such a j and any $\boldsymbol{w} \in \mathcal{W}$. Write $s_i = w_i/\Delta_i$ and $d_i^+ = \mathrm{kl}(\mu_i, \mu_1 + \varepsilon)$. The arm–1 branch in (50) gives

$$\Psi_j(\boldsymbol{w}) \leq s_j \mathrm{kl}(\mu_j, \mu_1 - \varepsilon) \leq \ell s_j d_j^+.$$

For any $a \notin \{j, 1\}$, choose the feasible pair-branch alternative $\lambda_j = \mu_j$ and $\lambda_a = \mu_j + \varepsilon$. Since $\mu_j + \varepsilon \leq \mu_1 + \varepsilon$ and $\mu_j + \varepsilon \geq \mu_1 \geq \mu_a$, monotonicity of $y \mapsto \mathrm{kl}(\mu_a, y)$ for $y \geq \mu_a$ gives

$$\Psi_j(\boldsymbol{w}) \leq s_a \mathrm{kl}(\mu_a, \mu_j + \varepsilon) \leq s_a d_a^+.$$

Combining these inequalities,

$$\Psi_j(\boldsymbol{w}) \leq \ell \min_{i \geq 2} s_i d_i^+.$$

Taking the supremum over $\boldsymbol{w} \in \mathcal{W}$ yields

$$\frac{1}{C_j(\boldsymbol{\mu})} \leq \ell \sup_{\boldsymbol{w} \in \mathcal{W}} \min_{i \geq 2} \frac{w_i}{\Delta_i} \mathrm{kl}(\mu_i, \mu_1 + \varepsilon) = \frac{\ell}{C^*(\boldsymbol{\mu})},$$

where the last equality is Proposition 4.3. Thus $C^*(\boldsymbol{\mu}) \leq \ell C_j(\boldsymbol{\mu})$ for every $j \in \mathcal{I}_\varepsilon(\boldsymbol{\mu})$. Taking the minimum over j gives $C^*(\boldsymbol{\mu}) \leq \ell C(\boldsymbol{\mu})$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: The abstract and introduction state the paper's main contributions: regret-aware fixed-confidence ε-BAI, instance-dependent regret lower bounds for ε-PAC and ε-PAC* policy families, the `KLUCB+Greedy` algorithm, and its regret/sample-complexity guarantees. These claims match the formal results in Sections 4–5 and the experiments in Appendix A.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper states the one-dimensional exponential-family assumptions in Section 4.1, discusses the gap between $C(\boldsymbol{\mu})$ and $C^*(\boldsymbol{\mu})$ in Section 4.3, and identifies open directions in Section 6, including whether one can design an algorithm with asymptotically optimal regret among all ε-PAC policies.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that are not acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper is primarily theoretical. The model assumptions are stated in Section 4.1, the main results are stated in Sections 4–5, and full proofs are provided in the appendices, including the lower-bound proofs, martingale/change-of-measure arguments, correctness proof, regret bound, sample-complexity analysis, and characteristic-time arguments.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix A specifies the simulated instances, reward families, value of ε, number of arms, number of repetitions, baselines, stopping rule used for KL-UCB, and reported metrics, which are the information needed to reproduce the synthetic experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), release of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example:
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The experiments use synthetic bandit instances whose parameters are disclosed, so no external dataset is required. However, the current manuscript does not provide an open-source code repository or executable scripts for reproducing the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiments are synthetic bandit simulations rather than train/test learning tasks. Appendix A reports the bandit instances, reward distributions, ε, the algorithms compared, the stopping rule modification for KL-UCB, the number of independent trials, and the metrics plotted.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental figures report averages over 500 independent repetitions, and the regret plots include error bars denoting one standard deviation above/below the mean. The paper also states the number of repetitions used for the sample-complexity plots.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.).
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar rather than state that they have a 96% confidence interval, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have stated in the experiments section that all experiments were implemented in MATLAB and parallelized on an Intel® Xeon® Gold 6244 CPU at 3.60 GHz with 32 GB of memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that did not make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and uses only synthetic simulations. It does not involve human subjects, private data, personally identifiable information, or high-risk model/data releases.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The introduction and Impact Statement discuss the positive clinical-trial motivation, where regret reduction may reduce exposure to inferior treatments.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of machine learning).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data or models that pose a high risk for misuse; it studies a theoretical bandit algorithm and synthetic simulations.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use external datasets, pretrained models, or third-party code assets as part of the reported experiments. Prior algorithms and theoretical results used for comparison are credited through citations.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `https://paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, benchmark, pretrained model, or software asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, human-subject experiments, or data collection from human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology, theory, algorithm, and experiments do not use LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.