Regret Tail Characterization of Optimal Bandit Algorithms with Generic Rewards

Subhodip Panda
Department of ECE
Indian Instiute of Science
Bangalore, India
subhodipp@iisc.ac.in

Shubhada Agrawal
Department of ECE
Indian Instiute of Science
Bangalore, India
shubhada@iisc.ac.in

***Abstract*—We consider the regret minimization problem in a stochastic multi-armed bandit setup, where the classical goal is to design policies that minimize the so-called expected cumulative regret. While expected-regret guarantees are well understood, controlling the tail behavior of the regret is crucial for their use in safety-critical applications. Recent works highlight that certain optimal algorithms can be fragile in the sense that they may incur large regret with non-negligible probability. However, current analysis has so far been confined mainly to simpler settings within the single parametric exponential family of reward distributions. Thus, the principal aim of our work is to analyze such regret-tail behavior of optimal bandit algorithms in a relatively broader setting: policies that are optimized for generic families of reward distributions under significantly weaker structural assumptions. For such general classes of reward distributions, we first show that a generalization of $\mathrm{KL}_{\inf}$-UCB algorithm remains asymptotically optimal. We then analyze its regret tail behavior and show that, for the distribution class under consideration, optimal algorithms generally exhibit a weak fragility in general. However, when the additional condition of *discrimination equivalence* holds, this fragility intensifies to a *strong* form characterized by heavy (Cauchy-type) regret tails. Since several widely studied model classes, including moment-bounded and bounded-support distributions, are contained in our framework, these results imply that optimal bandit algorithms for such families are strongly fragile whenever discrimination equivalence is satisfied. Finally, in the absence of discrimination equivalence, we refine the regret-tail upper-bound analysis and establish that, for bounded and finitely supported distributions, the optimal algorithm exhibits only a strictly weaker, *near-robust* form of fragility.**

I. Introduction

The multi-armed bandit (MAB) problem is a classical framework for sequential decision-making under uncertainty, in which an agent (player) interacts with a set of K arms, each associated with an unknown reward distribution. At each round, the agent selects an arm and observes a stochastic reward, with the goal of maximizing the total expected reward till time $T > 0$. This objective is typically formalized via the notion of regret, defined as the difference between the cumulative reward of an agent's policy over a horizon T and that of a policy that always plays an arm with the largest mean reward. Minimizing *expected* regret is equivalent to maximizing expected cumulative reward, and serves as the standard performance criterion in bandit problems.

A dominant focus of the contemporary literature has been on designing algorithms that minimize the expected regret. [1], [2] proposed a fundamental lower bound on the expected cumulative regret suffered by any reasonable algorithm for this setup. Algorithms achieving this lower bound (even in the multiplicative constants) have also been developed for a wide class of reward distributions [3]–[12]. We refer the reader to [13], [14] for a comprehensive treatment of this classical setup.

While expected regret is a canonical performance metric in stochastic bandits, it provides only a coarse summary of algorithmic behavior. A more informative characterization is obtained by studying the distribution of the regret [15]–[18], and in particular its tail behavior. Understanding the upper tail of the regret distribution is crucial for quantifying the probability of rare but severe failure events in which an algorithm may incur considerable regret. Such high-regret outcomes can be especially consequential in applications such as clinical trials, where such events may correspond to exposing a large number of patients to suboptimal treatments. Consequently, controlling heavy regret tails, rather than merely minimizing expected regret, is a key objective in risk-sensitive and safety-critical settings. A theoretical understanding of regret tail probabilities can thus provide a principled avenue for identifying and mitigating the vulnerabilities of existing optimal algorithms, and for guiding the design of more robust bandit strategies (see also, [19]).

In recent work, [20] establish a lower bound on the tail probabilities of the regret for any asymptotically optimal bandit algorithm, that is, one which matches the leading-order term in the regret lower bound, including the exact multiplicative constant. They further provide a detailed analysis of the KL-UCB algorithm of [9], showing that its regret tail matches their lower bound, thereby demonstrating the tightness of these bounds for parametric settings such as single-parameter exponential family (SPEF) models. While the authors discuss the extensions of their lower bounds to certain non-parametric classes of reward distributions, the regret tail behavior of optimized algorithms in these more general settings remains open.

In particular, their analysis does not extend to the popular empirical KL-UCB, which is known to be optimal for both finitely-supported distributions (Theorem 2 in [9]) and bounded

supported distributions (Proposition 4 in [12]), nor to the more general $\mathrm{KL}_{\inf}$-UCB algorithm proposed in [12]. This motivates the central question of the present work:

Can we characterize the regret tail behavior for asymptotically optimal bandit algorithms for a broader, nonparametric class of reward distributions?

In this work, we address the above question in affirmative. We first show that the generalization of KL-UCB algorithm, studied in [12] for a specific heavy-tailed family ($\mathrm{KL}_{\inf}$-UCB), is actually asymptotically optimal for a very broad class of distributions $\mathcal{L}$, introduced later. We then prove a regret-tail upper bound for the $\mathrm{KL}_{\inf}$-UCB algorithm for this general family of arm distributions. We conclude this section by presenting the key contributions of this work.

- We extend the $\mathrm{KL}_{\inf}$-UCB algorithm of [12] to a much broader family of reward distributions $\mathcal{L}$ (to be introduced later), and show that it is asymptotically optimal (Theorem III.1). This is a substantial generalization, and includes the moment-bounded class studied by the authors, as well as the bounded support distributions as special cases.
- We further prove a regret tail upper bound for this generalized $\mathrm{KL}_{\inf}$-UCB algorithm (Theorem IV.5). As immediate corollaries, we the upper bounds for the optimal UCB algorithms in classical settings: *(a)* $\mathrm{KL}_{\inf}$-UCB for moment-bounded class (Corollary V.2); *(b)* empirical KL-UCB for bounded-support reward distributions (Corollary V.3). When the class $\mathcal{L}$ satisfies an additional property called discrimination-equivalence (Definition IV.3), we can show that the upper bounds that we prove exactly match the regret tail lower bound proposed in [20] (Remark IV.6).
- For the special case of finitely-supported distributions, which do not satisfy the discrimination-equivalence property, we provide a much tighter regret tail upper bound for the empirical KL-UCB algorithm (Theorem VI.1), and show that it matches the lower bound of [20].

Organization: The rest of the paper is organized as follows. Section II presents the setup and necessary background. In Section III, we present the *Generalized* $\mathrm{KL}_{\inf}$-UCB algorithm and analyze its asymptotic optimality. Section IV analyzes the regret tail behavior of this $\mathrm{KL}_{\inf}$-UCB under the discrimination-equivalence condition. Finally, Section VI provides a refined regret tail analysis for empirical KL-UCB in the finitely supported setting to obtain tight upper bounds without relying on discrimination equivalence. We review the relevant literature in section Appendix B.

II. Setup and preliminaries

We consider a bandit problem with K-arms indexed by $a \in [K] := \{1, \dots, K\}$, with $K \geq 2$. Let $\mathcal{P}(\Re)$ denote the collection of all probability measures on $\Re$. For $\nu \in \mathcal{P}(\Re)$, let $m(\nu) = \int_{\Re} x d\nu$ denote the mean of distribution ν. Let $\mathcal{M} \subseteq \mathcal{P}(\Re)$ be any collection of probability measures. We call a bandit model $\mathcal{M}^K$, which denotes the collection of vectors of K distributions, each from $\mathcal{M}$. Now, given a bandit environment $\boldsymbol{\nu} \in \mathcal{M}^K$ such that $\boldsymbol{\nu} = \{\nu_1, \nu_2, \dots, \nu_K\}$, each arm $a \in [K]$ has a reward distribution ν_a with expected reward $\mu_a = m(\nu_a)$. Let $\mu^* = \max\{\mu_a : a \in [K]\}$ be the highest expected reward associated with the optimal arm. We denote the sub-optimality gap of an arm a as $\Delta_a = \mu^* - \mu_a$. Without the loss of generality, we assume that arm-1 is the optimal arm such that $\mu_1 > \mu_2 > \ldots > \mu_K$ for the rest of the paper. Now, at time t, the agent (player) selects an arm $A_t \in [K]$ based on the past information to receive a reward Y_t. Now, the number of times each arm a is pulled till time t is referred to as $N_a(t) \triangleq \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$. In addition, for each arm a and all rounds t such that $N_a(t) \geq 1$, the empirical reward distribution of arm-a is defined as $\hat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{I}\{A_s = a\}$. Here δ_{Y_s} denotes the Dirac measure at Y_s. The quality of an algorithm (policy) π is evaluated using the standard notion of *expected regret*, which we now define formally. The expected regret (or simply pseudo-regret) at round $T \geq 1$ is defined as follows.

$$\mathbb{E}[R(T)] \triangleq \mathbb{E}\left[T\mu_1 - \sum_{t=1}^{T} Y_t\right] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)].$$

Note that the above expectation is with respect to the probability measure $\mathbb{P}_{\boldsymbol{\nu}}^{\pi}$ which is induced by the interaction between the algorithm π and the bandit environment $\boldsymbol{\nu}$. Now we define what it means for an algorithm to be called optimal for a bandit model using the Lai-Robbins lower bound. Following the seminal work of [1], the minimal achievable growth rate of the expected regret for algorithms designed for the model $\mathcal{M}^K$ is precisely characterized. Further [2] generalized this notion of Lai-Robbins lower bound, which is used by [20] for defining an optimal algorithm as follows.

Definition II.1 (Optimal Algorithm)**.** An algorithm is $\mathcal{M}^K$-optimal algorithm if for any environment $\boldsymbol{\nu} \in \mathcal{M}^K$ and for each sub-optimal arm a, the following holds

$$\lim_{T \to \infty} \frac{\mathbb{E}[N_a(T)]}{\log T} = \frac{1}{\mathrm{KL}_{\inf}^{\mathcal{M}}(\nu_a, \mu_1)}.$$

The proof of the Lai-Robbins lower bound [1], [2] relies on change of measure arguments (see [13]). [21, Lemma 1] show that it is necessary to impose certain restrictions on the class $\mathcal{M}$ under consideration, otherwise $\mathrm{KL}_{\inf}(\cdot, \cdot) = 0$ leading to unbounded expected regret. To this end, we restrict $\mathcal{M}$ to the class $\mathcal{L}$ having the $\mathrm{KL}_{\inf}$-concentration properties as defined in Assumption II.2 and Assumption II.3 below. Finally, in the subsequent section, we present a generalized version of $\mathrm{KL}_{\inf}$-UCB algorithm [12], which is asymptotically optimal (Theorem III.1) for the distribution class $\mathcal{L}$.

A. Distribution Class $\mathcal{L}$

Recall from [22] that without any restrictions on the class of reward distributions, the lower bound on the expected regret becomes unbounded (logarithmic regret is impossible). We therefore restrict our attention to the so-called $\mathrm{KL}_{\inf}$-concentrated class of distributions, which we denote by $\mathcal{L}$.

To specify $\mathcal{L}$, we need certain definitions, which we now introduce.

For a probability distribution ν and $x \in \Re$, $\mathrm{KL}_{\inf}^{\mathcal{L}}(\nu, x)$, defined below, has two concentration properties given by Assumption II.2 and Assumption II.3. We define the $\mathrm{KL}_{\inf}^{\mathcal{L}}(\nu, x)$ for a distribution ν and $x \in \mathbb{R}$ as follows:

$$\mathrm{KL}_{\inf}^{\mathcal{L}}(\nu, x) = \inf\{\mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{L} \text{ and } \mathbb{E}[\nu'] \geq x\}$$

Here $\mathrm{KL}(\nu, \nu') = \int \log \frac{d\nu}{d\nu'} d\nu$ denotes the Kullback-Leibler divergence between two distributions ν and ν'. For notational simplicity, we drop the superscript $\mathcal{L}$ to denote $\mathrm{KL}_{\inf}(\nu, x)$. Now, we discuss the two concentration properties of $\mathrm{KL}_{\inf}(\nu, x)$ that characterize the distribution class $\mathcal{L}$ as follows.

Assumption II.2. Let $\hat{\nu}_n$ be the empirical distribution of ν having mean $m(\nu)$ and $g(n)$ be an increasing function such that $g(n) = O(\log(1+n))$, then the following holds:

$$\mathbb{P}\left(\exists n \in \mathbb{N} : n \cdot \mathrm{KL}_{\inf}(\hat{\nu}_n, m(\nu)) - g(n) \geq x\right) \leq e^{-x} \quad (1)$$

Assumption II.3. Let $\hat{\nu}_n$ be the empirical distribution of ν having mean $m(\nu)$ and $\delta > 0$,then there exists constants $d_0 > 0$ and $c_\nu > 0$ s.t. for all $d < d_0$ the following holds.

$$\mathbb{P}\Big(\mathrm{KL}_{\inf}(\hat{\nu}_n, m(\nu)+\delta) \leq \mathrm{KL}_{\inf}(\nu, m(\nu)+\delta) - d\Big) \leq e^{-nc_\nu d^2} \quad (2)$$

These two assumed concentration in equations 1 and 2 for $\mathrm{KL}_{\inf}(\nu, x)$ provide structural restrictions for the class of distributions $\mathcal{L}$. This class of distribution is fairly broad because various useful families of distributions, such as bounded support [3] and moment-bounded distributions [12], follow these assumptions. We discuss these examples in detail in Appendix A.1.

III. Generalized $\mathrm{KL}_{\inf}$-UCB Algorithm and its Optimality

In this section, we present a straight-forward generalization of the $\mathrm{KL}_{\inf}$-UCB algorithm for arm distributions from $\mathcal{L}$, and prove that it is asymptotically optimal. Let $\boldsymbol{\nu} \in \mathcal{L}^K$ denote a K-armed bandit instance.

For exploration functions $f_a(\cdot)$ for each $a \in [K]$, define the index of arm a as

$$U_a(N_a(t), t) = \sup\left\{x \in \Re : \mathrm{KL}_{\inf}(\hat{\nu}_a(t), x) \leq \frac{f_a(t)}{N_a(t)}\right\}.$$

The algorithm initializes by pulling each arm once. At each subsequent time instance t, it computes the index $U_a(N_a(t), t)$ for every arm, and pulls the arm with the maximum value of the computed index (ties broken arbitrarily).

We remark that the prior work [12] introduced a batched variant of the algorithm to reduce the computational overhead of the naive approach. However, this batched procedure is not asymptotically optimal (its regret matches the lower bound only up to a multiplicative constant multiplicative). Hence, we restrict our attention to the single-batch version, and show that it is asymptotically optimal for $\mathcal{L}$.

Algorithm 1 *Generalized* $\mathrm{KL}_{\inf}$-UCB$(K, \{f_a(\cdot)\}_{a=1}^K)$

Input: K; $\mathcal{L}$; exploration functions for each arm, i.e., $f_a(\cdot)$.
Initialization: Pull each arm $a \in [K]$ once
Set $t \longleftarrow K+1$
Compute $\hat{\nu}_a(t)$, and update $N_a(t)$ for all arms $a \in [K]$.

for $t = K+1$ to T **do**
  for each arm $a \in \{1, \ldots, K\}$ **do**
    Compute index $U_a(N_a(t), t) = \sup\left\{\mathbb{E}(\nu) : \nu \in \mathcal{L}, \mathrm{KL}(\hat{\nu}_a(t), \nu) \leq \frac{f_a(t)}{N_a(t)}\right\}$
  end for
  Pick an arm $A_{t+1} \in \arg\max_{a \in [K]} U_a(N_a(t), t)$
  Set $t \leftarrow t+1$
  Update $\hat{\nu}_a(t)$, and update $N_a(t)$ for all arms $a \in [K]$.
end for

Theorem III.1. *For $\boldsymbol{\nu} \in \mathcal{L}^K$, and $f_a(t) = \log(t) + 2\log\log(t) + 2\log(1+N_a(t)) + 1$, Generalized $\mathrm{KL}_{\inf}$-UCB, with inputs $(K, f_a(.))$ is asymptotically optimal, i.e.,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \leq \frac{1}{\mathrm{KL}_{\inf}(\nu_a, \mu_1)}.$$

The proof of the above theorem closely follows that in [12, Theorem 1] and is presented in the Appendix A.2. In the following section, we now formally define what it means to say an algorithm is fragile and show one of our key results for the *Generalized* $\mathrm{KL}_{\inf}$-UCB.

IV. Regret Tail Fragility in Optimal Algorithms

[20] proposed a key central result that lower bounds the high regret tail events. This lower bound indicates that large-regret events occur with some non-negligible probability, thereby demonstrating the fragility of the algorithm. We formally present this lower bound proposed by [20] in the below Theorem IV.1 as follows.

Theorem IV.1 (Optimal Regret Tail Lower-bound). *Let π be $\mathcal{M}^K$-optimal algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{M}^K$, a deviation family $\mathcal{D}_\gamma(T) = \left[\log^{1+\gamma}(T), (1-\gamma)T\right]$, such that $\gamma \in (0,1)$, and for an i-th best arm,*

$$\liminf_{T\to\infty} \inf_{x \in \mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}_\nu^\pi(N_i(T) > x)}{\log x} \geq -\sum_{j=1}^{i-1} \inf_{\substack{\widetilde{\nu} \in \mathcal{M}: \\ m(\widetilde{\nu}) < \mu_i}} \frac{\mathrm{KL}(\widetilde{\nu}, \nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu}, \mu_i)}. \quad (3)$$

The proof of the above theorem is provided in Section 3.2 of [20] for the special case where $\mathcal{M}^K$ is an SPEF model. Closely following this, the proof for general model classes is given in Appendix A.3.

Now, note that each term inside the summation on the right-hand side of the equation (3) denoted as $C_{\nu_j} = \frac{\mathrm{KL}(\widetilde{\nu}, \nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu}, \mu_i)} \geq 1$.

This directly follows from the definition of $\mathrm{KL}_{\inf}(\cdot,\cdot)$ with $\mu_j > \mu_i$. For the special case when $C_{\nu_j} = 1$, the regret tail event $\mathbb{P}(R(T) > x)$ exhibits polynomial decay of order x^{-1} on a logarithmic scale over a wide range of deviations. This behavior matches a heavy-tailed Cauchy-type distribution truncated at the time horizon T. We call this special case *Cauchy Fragility*, which is defined as follows.

Definition IV.2 (Cauchy Fragility). Let π be $\mathcal{M}^K$-optimal algorithm. We say that π achieves *Cauchy Fragility* if for any environment $\boldsymbol{\nu} \in \mathcal{M}^K$, deviation family $\mathcal{D}$ the following holds,

$$\lim_{T\to\infty} \inf_{x\in\mathcal{D}} \frac{\log \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(N_i(T) > x)}{\log x} = -(i-1)$$

It is important to emphasize that *Cauchy Fragility* represents a stronger notion of fragility reflecting truly heavy-tailed regret behavior. A natural question that arises is whether optimal algorithms necessarily exhibit this stronger notion of *Cauchy Fragility*. As noted earlier, this regime emerges only when $C_{\nu_j} = 1$, a condition that holds when the underlying model class $\mathcal{M}$ satisfies the property of discrimination equivalence, which we formally define below.

Definition IV.3 (Discrimination Equivalance). We say a distribution class $\mathcal{M}$ is discrimination equivalent if for any distributions $\nu, \nu' \in \mathcal{M}$ such that $m(\nu) > m(\nu')$ the following holds

$$\inf_{\substack{\widetilde{\nu}\in\mathcal{M}:\\ m(\widetilde{\nu})<m(\nu')}} \frac{\mathrm{KL}(\widetilde{\nu},\nu)}{\mathrm{KL}_{\inf}(\widetilde{\nu}, m(\nu'))} = 1$$

The above definition generalizes the notion of discrimination equivalence introduced in the previous work (See Definition 3 in [20]) for SPEF models. Note that, in the absence of discrimination equivalence, some optimal algorithms exhibit only a relaxed form of fragility, characterized by substantially lighter regret tails. As this is a comparatively less catastrophic form of fragility and matches the lower bound in (3), we refer to this as *Near-Robust Fragility*, which is defined as follows.

Definition IV.4 (Near-Robust Fragility). Let π be $\mathcal{M}^K$-optimal algorithm. We say that π achieves *Near-Robust Fragility* if for any environment $\boldsymbol{\nu} \in \mathcal{M}^K$, deviation family $\mathcal{D}$ the following holds,

$$\lim_{T\to\infty} \inf_{x\in\mathcal{D}} \frac{\log \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(N_i(T) > x)}{\log x} = -\sum_{j=1}^{i-1} \inf_{\substack{\widetilde{\nu}\in\mathcal{M}:\\ m(\widetilde{\nu})<\mu_i}} \frac{\mathrm{KL}(\widetilde{\nu},\nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu},\mu_i)}$$

In particular, [20] established that without discrimination equivalence, the KL-UCB algorithm [9], which is asymptotically optimal for SPEF bandit models, achieves *Near-Robust Fragility*. However, their analysis is inherently restricted to exponential family models and does not extend to broader distribution classes. For instance, it does not apply to the empirical KL-UCB algorithm, which is known to be asymptotically optimal for bounded, finitely supported distributions [9] as well as for heavy-tailed distributions [12]. In order to analyze the fragility issues of optimal bandit algorithms beyond exponential families, we present our key result in the following theorem.

Theorem IV.5 (Regret Tail Upper Bound for distribution class $\mathcal{L}$). *Let π be $\mathcal{L}^K$-optimal Generalizeded $\mathrm{KL}_{\inf}$-UCB algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{L}^K$, a deviation family $\mathcal{D}_\gamma(T) = \big[\log^{1+\gamma}(T),\, (1-\gamma)T\big]$, such that $\gamma \in (0,1)$, and for an i-th best arm,*

$$\limsup_{T\to\infty} \inf_{x\in\mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(N_i(T) > x)}{\log x} \;\le\; -(i-1) \qquad (4)$$

We defer the proof of Theorem IV.5 to Appendix A.4. Combining Theorems IV.1 and IV.5, we observe that for the broader distribution class $\mathcal{L}$, there exists a nontrivial gap between the regret tail lower bound and the corresponding upper bound. Consequently, this gap implies a comparatively weaker form of fragility as the regret tail is lighter than that of the Cauchy distribution. In the following sections, we provide further insights on the Cauchy fragility and the Near-Robust fragility.

Remark IV.6. When the class $\mathcal{L}$ satisfies the discrimination equivalence property, the regret tail lower bound in (3) simplifies to $-(i-1)$, which exactly matches the regret tail upper bound in (4).

V. Cauchy Fragility of Optimal Algorithms

From Remark IV.6, we see that in discrimination equivalent classes, this optimal algorithm is *Cacuchy Fragile*, as formalized in the following corollary.

Corollary V.1. *Let π be $\mathcal{L}^K$-optimal Generalizeded $\mathrm{KL}_{\inf}$-UCB algorithm. If $\mathcal{L}$ is discrimination equivalant then, for any environment $\boldsymbol{\nu} \in \mathcal{L}^K$, a deviation family $\mathcal{D}_\gamma(T) = \big[\log^{1+\gamma}(T),\, (1-\gamma)T\big]$, such that $\gamma \in (0,1)$, π achieves Cauchy fragility.*

The proof of Corollary V.1 follows directly from the definitions. In particular, substituting the discrimination equivalence condition into the regret tail lower bound in (3), and comparing it with the regret tail upper bound in (4) from Theorem IV.5, yields the desired result. We emphasize that this result has two important implications arising from the generality of the distribution class $\mathcal{L}$ and the flexibility of the $\mathrm{KL}_{\inf}$-UCB framework. We elaborate on these implications as follows.

Fragility of $\mathrm{KL}_{\inf}$-UCB under Moment-Bounded Models: We consider the class of reward distributions whose $(1+\varepsilon)$-th moments are uniformly bounded. For $\varepsilon > 0$ and $B > 0$, define $\mathcal{L}_{B,\varepsilon} \triangleq \big\{\nu \in \mathcal{P}(\mathbb{R}) :\, \mathbb{E}_\nu\big[|X|^{1+\varepsilon}\big] \le B\big\}$. We know, every distribution $\nu \in \mathcal{L}_{B,\varepsilon}$ satisfies the structural conditions imposed on the distribution class $\mathcal{L}$ in Assumptions II.2 and II.3. Since the *Generalized* $\mathrm{KL}_{\inf}$-UCB algorithm is asymptotically optimal over $\mathcal{L}_{B,\varepsilon}$, the general fragility results established for the broader class $\mathcal{L}$ immediately apply. This yields the following corollary,

Corollary V.2. *Let π be $\mathcal{L}^K_{B,\varepsilon}$-optimal Generalized $\mathrm{KL}_{\inf}$-UCB algorithm. If $\mathcal{L}_{B,\varepsilon}$ is discrimination equivalant then, for any environment $\boldsymbol{\nu} \in \mathcal{L}^K_{B,\varepsilon}$, a deviation family $\mathcal{D}_\gamma(T) =$*

$\left[\log^{1+\gamma}(T),\,(1-\gamma)T\right]$, *such that* $\gamma \in (0,1)$, π *achieves Cauchy fragility.*

Fragility under Bounded-Support Models: We now consider the class of reward distributions with bounded support. For a constant $a, b \in \Re$, define $\mathcal{L}_{a,b} \triangleq \{\nu \in \mathcal{P}(\mathbb{R}) : \mathrm{Supp}(\nu) \subseteq [a,b]\}$. Similar to the above, every distribution $\nu \in \mathcal{L}_{a,b}$ satisfies the structural assumptions defining the class $\mathcal{L}$ (see Assumptions II.2 and II.3). As discussed in Section III, the *Generalized* $\mathrm{KL}_{\inf}$-UCB framework specializes to the empirical KL-UCB algorithm of [9] under an appropriate choice of the exploration function. In particular, empirical KL-UCB corresponds to the choice $f_a(t) = g(t) + h(a,t)$ with $g(t) = \log t + \log\log t$ and $h(a,t) \equiv 0$ for all arms a. Since empirical KL-UCB is asymptotically optimal for the bounded-support model $\mathcal{L}_{a,b}$, the fragility results established for the broader class $\mathcal{L}$ apply directly. This yields the following corollary,

Corollary V.3. *Let π be $\mathcal{L}_B^K$-optimal empirical KL-UCB algorithm. If $\mathcal{L}_B$ is discrimination equivalant then, for any environment $\boldsymbol{\nu} \in \mathcal{L}_B^K$, a deviation family $\mathcal{D}_\gamma(T) = \left[\log^{1+\gamma}(T),\,(1-\gamma)T\right]$, such that $\gamma \in (0,1)$, π achieves Cauchy fragility.*

Corollary V.2 follows immediately from Corollary V.1 by a direct specialization of the distribution class under consideration. In contrast, the proof of Corollary V.3 doesn't directly follow because of the exploration function being different from *Generalized* $\mathrm{KL}_{\inf}$-UCB algorithm. We present the detailed proof of Corollary V.3 in Appendix A.5.

VI. Near-Robust Fragility of Optimal Algorithms

In the preceding section, we characterized the regret tail (fragility) behavior under the additional assumption of *discrimination equivalence*. While this condition highlights a stronger fragility issue, it can be restrictive in practice. This naturally motivates the following question: *Do optimal algorithms exhibit near-robust fragility in the absence of discrimination equivalence?* Current results show a non-trivial gap between the lower and upper bounds for the broad distribution class $\mathcal{L}$. Nevertheless, for a more structured subclass, bounded and finitely supported distributions, we are able to substantially refine our analysis. Formally, let $\mathcal{L}_{a,b,s}$ denote the class of distributions supported on at most $s < \infty$ points contained in the interval $[a,b]$, i.e., $\mathcal{L}_{a,b,s} \triangleq \{\nu \in \mathcal{P}(\Re) : \mathrm{Supp}(\nu) \subseteq [-a,b],\ |\mathrm{Supp}(\nu)| \le s\}$. In particular, we show that the empirical KL-UCB algorithm, which is known to be asymptotically optimal for this model class $\mathcal{L}_{a,b,s}$, actually achieves *Near-Robust Fragility* as formally stated in Theorem VI.1 as follows,

Theorem VI.1 (Near-Robust Fragility of empirical KL-UCB). *Let π be $\mathcal{L}_{a,b,s}^K$-optimal empirical KL-UCB algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{L}_{B,s}^K$, a deviation family $\mathcal{D}_\gamma(T) = \left[\log^{1+\gamma}(T),\,(1-\gamma)T\right]$, such that $\gamma \in (0,1)$, and for an i-th best arm,*

$$\lim_{T\to\infty} \inf_{x\in\mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}_{\boldsymbol{\nu}}^{\pi}(N_i(T) > x)}{\log x} = -\sum_{j=1}^{i-1} \inf_{\substack{\widetilde{\nu}\in\mathcal{M}:\\ m(\widetilde{\nu})<\mu_i}} \frac{\mathrm{KL}(\widetilde{\nu},\nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu},\mu_i)}$$

Proof Outline: The proof proceeds by decomposing the regret-tail event $\mathbb{P}_{\boldsymbol{\nu}}^{\pi}(N_i(T) > x)$ into two complementary event components. The first component is controlled by partitioning the event, according to the number of pulls of sub-optimal arms, and applying a union bound over the resulting sub-events. Each such subevent is then bounded using a finite-sample version of Sanov's theorem [23], [24], which yields sharp exponential deviation bounds for empirical measures. The second component is controlled using the concentration properties of the $\mathrm{KL}_{\inf}$ functional established in Assumption II.3. Combining these bounds yields the desired upper bound on the regret tail. A complete provided in Appendix A.6.

VII. Conclusion, Limitations and Future Works

In this work, we investigated the regret-tail behavior of stochastic multi-armed bandit algorithms that are optimal over a broad class of reward distributions. Our results significantly extend earlier fragility phenomena previously established primarily for exponential-family models to much broader distribution classes, including moment-bounded and bounded-support distributions. We show that, in general, optimal algorithms for this broader class exhibit a relatively weaker form of fragility, in the sense that their regret tails are lighter than those of a Cauchy distribution. However, under the additional condition of discrimination equivalence, optimal algorithms necessarily display strong Cauchy fragility: the regret tail becomes heavy, implying that large-regret events occur with non-negligible probability. Furthermore, in the absence of discrimination equivalence, we show that optimal algorithms for bounded and finitely supported reward models exhibit a strictly weaker form of near-robust fragility, wherein the regret-tail upper bound exactly matches the optimal regret-tail lower bound. Despite these advances, several limitations remain. For the general distribution class $\mathcal{L}$, a gap persists between the regret-tail upper and lower bounds when discrimination equivalence does not hold. Closing this gap and designing algorithms that achieve improved tail robustness constitute important directions for future work. Further, the study of fragility issues in the Thompson sampling [6], [25] remains a promising direction.

References

[1] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4 – 22, 1985.

[2] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for sequential allocation problems," *Advances in Applied Mathematics*, vol. 17, no. 2, pp. 122 – 142, 1996.

[3] J. Honda and A. Takemura, "An asymptotically optimal bandit algorithm for bounded support models," in *In Proceedings of the Twenty-third Conference on Learning Theory (COLT 2010*. Omnipress, 2010, pp. 67–79.

[4] H. Junya and T. Akimichi, "An asymptotically optimal policy for finite support models in the multiarmed bandit problem," *Machine Learning*, vol. 85, no. 3, pp. 361–391, 2011.

[5] J. Honda and A. Takemura, "An asymptotically optimal policy for finite support models in the multiarmed bandit problem," *Machine Learning*, vol. 85, no. 3, pp. 361–391, Dec 2011.

[6] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 2012, pp. 39–1.

[7] A. Shipra and G. Navin, "Further optimal regret bounds for thompson sampling," in *Artificial intelligence and statistics*. PMLR, 2013, pp. 99–107.

[8] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *International Conference on Algorithmic Learning Theory*. Springer, 2012, pp. 199–213.

[9] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz *et al.*, "Kullback–Leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, vol. 41, no. 3, pp. 1516–1541, 2013.

[10] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tails," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7711–7717, 2013.

[11] J. Honda and A. Takemura, "Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3721–3756, 2015.

[12] S. Agrawal, S. K. Juneja, and W. M. Koolen, "Regret minimization in heavy-tailed bandits," in *Proceedings of the Thirty Fourth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, M. Belkin and S. Kpotufe, Eds., vol. 134. PMLR, 2021, pp. 26–62. [Online]. Available: https://proceedings.mlr.press/v134/agrawal21a.html

[13] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[14] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[15] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi–armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[16] A. Salomon and J.-Y. Audibert, "Deviations of stochastic bandit regret," in *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2011, pp. 159–173.

[17] A. Kalvit and A. Zeevi, "A closer look at the worst-case behavior of multi-armed bandit algorithms," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[18] L. Fan and P. W. Glynn, "The typical behavior of bandit algorithms," *CoRR*, vol. abs/2210.05660, 2022. [Online]. Available: https://arxiv.org/abs/2210.05660

[19] K. Ashutosh, J. Nair, A. Kagrecha, and K. Jagannathan, "Bandit algorithms: Letting go of logarithmic regret for statistical robustness," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, 2021.

[20] L. Fan and P. W. Glynn, "The fragility of optimized bandit algorithms," *Operations Research*, 2024.

[21] S. Agrawal, S. Juneja, and P. Glynn, "Optimal δ-correct best-arm selection for heavy-tailed distributions," in *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 117. PMLR, 08 Feb–11 Feb 2020, pp. 61–110.

[22] A. Shubhada, J. Sandeep, and G. Peter, "Optimal δ-correct best-arm selection for heavy-tailed distributions," in *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 117. PMLR, 2020, pp. 61–110.

[23] A. Dembo and O. Zeitouni, "Large deviations techniques and applications. corrected reprint of the second (1998) edition. stochastic modelling and applied probability, 38," 2010.

[24] I. Csiszár, "A simple proof of sanov's theorem," *Bulletin of the Brazilian Mathematical Society, New Series*, vol. 37, no. 3, pp. 453–459, 2006. [Online]. Available: https://doi.org/10.1007/s00574-006-0021-2

[25] C. Riou and J. Honda, "Bandit algorithms based on thompson sampling for bounded reward distributions," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 777–826.

[26] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[27] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 09 1952.

[28] T. L. Lai, "Adaptive treatment allocation and the multi–armed bandit problem," *The Annals of Statistics*, vol. 15, no. 3, pp. 1091–1114, 1987.

[29] R. Agrawal, "Sample mean based index policies with o (log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, pp. 1054–1078, 1995.

[30] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, May 2002.

APPENDIX

A.1 Generality of Distribution Class $\mathcal{L}$

In the introduction section, the properties of the distribution class $\mathcal{L}$ are defined in Assumption II.2 and Assumption II.3. Let's define this distribution class as follows.

$$\mathcal{L} = \{\nu \in \mathcal{P}(\Re) : \nu \text{ follows Assumption II.2 and Assumpiton II.3}\}$$

Now we provide two distribution classes that are examples of $\mathcal{L}$.

Moment Bounded Distributions: For $\varepsilon > 0$ and $B > 0$, define moment bounded distribution class as follows

$$\mathcal{L}_{B,\varepsilon} \triangleq \left\{\nu \in \mathcal{P}(\Re) : \mathbb{E}_\nu\left[|X|^{1+\varepsilon}\right] \leq B\right\}$$

This distribution class satisfies Assumptions II.2 and II.3. In particular, the validity of Assumption II.2 for the moment-bounded class of distributions is established in Proposition 5 of [12]. Similarly, Assumption II.3 follows directly from the arguments used in Lemma 6 of [12]. We refer the reader to the aforementioned results for a rigorous and complete proof.

Bounded Support Distributions: For constants $a, b \in \Re$, define bounded support distribution class

$$\mathcal{L}_{a,b} \triangleq \{\nu \in \mathcal{P}(\Re) : \text{Supp}(\nu) \subseteq [a, b]\}$$

This distribution class also satisfies the above assumptions. The proof follows along the same lines as those in [12], with an appropriate modification that employs the dual representation of KL_{inf} for bounded-support distributions as developed in Section 4 of [3].

A.2 Proof of Theorem III.1

This proof follows the argument of Theorem 1 in [12], specialized to the case where the batch size equals one in each trial. Take $t \geq K + 1$, and without loss of generality, assume that arm 1 is optimal. Then the event that, at the beginning of round t, a sub-optimal arm $a \neq 1$ attains the maximum index i.e., the event $\{A_t = a\}$ is contained in

$$\{U_1(N_a(t), t) \leq \mu_1 \quad \text{and} \quad A_t = a\} \bigcup \{U_a(N_a(t), t) > \mu_1 \quad \text{and} \quad A_t = a\} \tag{5}$$

The left-hand event of (5) characterizes the underestimation of the optimal arm's index relative to its true mean at time t, while the right-hand event corresponds to an overestimation of the sub-optimal arm's index beyond the mean of the optimal arm. Recall that during the initial K rounds, each arm is played exactly once as part of the initialization procedure. Thus,

$$N_a(T) = 1 + \sum_{t=K+1}^{T} \mathbb{I}\{A_t = a\}$$

$$\mathbb{E}[N_a(T)] \leq \ 1 + \mathbb{E}[D_T] + \mathbb{E}[E_T]$$

The terms D_T and E_T are as follows:

$$D_T := \sum_{t=K+1}^{T} \mathbb{I}\left(U_1(N_a(t), t) \leq \mu_1,\ A_t = a\right), \text{ and } E_T := \sum_{t=K+1}^{T} \mathbb{I}(U_a(N_a(t), t) > \mu_1,\ A_t = a).$$

Bounding the overestimation of sub-optimal arms in E_T: By the definition of the index employed by the algorithm, for any $t \geq K + 1$ and $x \in \mathbb{R}$, the event $\{U_a(N_a(t), t) \geq x\}$ is equivalent to $\{N_a(t)\,\text{KL}_{\text{inf}}(\hat{\nu}_a(t), x) \leq f_a(t)\}$. Let $d > 0$ satisfy $\min_{a>1} \text{KL}_{\text{inf}}(\nu_a, \mu_1) \geq d$. Then the indicator of the event $\{U_a(N_a(t), t) \geq \mu_1,\ A_t = a\}$ is dominated by the sum of the two events E_{1t} and E_{2t} defined below:

$$E_{1t} = \mathbb{I}\left(\text{KL}_{\text{inf}}(\hat{\nu}_a(t), \mu_1) \leq \frac{f_a(t)}{N_a(t)},\ \text{KL}_{\text{inf}}(\hat{\nu}_a(t), \mu_1) > \text{KL}_{\text{inf}}(\nu_a, \mu_1) - d,\ A_t = a\right)$$

$$E_{2t} = \mathbb{I}\left(\text{KL}_{\text{inf}}(\hat{\nu}_a(t), \mu_1) \leq \text{KL}_{\text{inf}}(\nu_a, \mu_1) - d,\ A_t = a\right).$$

Thus,

$$E_T \leq \sum_{t=K+1}^{T} E_{1t} + \sum_{t=K+1}^{T} E_{2t}$$

We can clearly see that E_{1t}, is bounded above by $\mathbb{I}\left(N_a(t)\left(\text{KL}_{\text{inf}}(\nu_a, \mu_1) - d\right) \leq f_a(t), A_t = a\right)$, giving

$$\sum_{t=K+1}^{T} E_{1t} \leq \sum_{t=1}^{T} E_{1t} \leq \sum_{t=1}^{T} \mathbb{I}\left(N_a(t) \leq \frac{f_a(t)}{\text{KL}_{\text{inf}}(\nu_a, \mu_1) - d}, A_t = a\right) \tag{6}$$

Now, in order to upper bound the RHS of equation (6), we use the following lemma from [12]. For completeness, we state the result as follows:

Lemma A.1. *For $T \geq K+1$, $\tilde{\eta} \geq 0$, $d > 0$,B_j be the size of the j^{th} batch and N be the number of batches till time T then*

$$\sum_{j=1}^{N} B_j E_{1j} \leq (1+\tilde{\eta}) \left(\frac{\log(T)}{\mathrm{KL}_{\inf}(\nu_a, \mu_1) - d} + O\left(\log\log(T)\right) \right).$$

We use the above Lemma A.1 with $\widetilde{\eta} = 0$ to essentially derive the following bounds

$$\sum_{t=1}^{T} E_{1t} \leq \frac{\log(T)}{\mathrm{KL}_{\inf}(\nu_a, \mu_1) - d} + O(\log\log(T)).$$

For the exact form of the $O(\log\log(T))$ term above, we refer the reader to look at the proof of Lemma 14 in [12].

Note that for any constant $c > 0$, $1 - e^{-c} \geq \frac{c}{1+c}$ because $e^c \geq 1 + c$. Now in order to control the events in E_{2t} we directly use AssumptionII.3 with $\delta = \mu_1 - \mu_a$ to get an upper bound as follows

$$\begin{aligned}
\mathbb{E}\left(\sum_{t=K+1}^{T} E_{2t} \right) &\leq \sum_{t=K+1}^{T} \mathbb{P}(\mathrm{KL}_{\inf}(\hat{\nu}_a(t), \mu_1) \leq \mathrm{KL}_{\inf}(\nu_a, \mu_1) - d) \\
&\leq \sum_{t=K+1}^{T} e^{-t c_{\nu_a} d^2} \\
&\leq \sum_{t=2}^{\infty} e^{-t c_{\nu_a} d^2} \\
&= \frac{e^{-2c_{\nu_a} d^2}}{1 - e^{c_{\nu_a} d^2}} \\
&\leq 1 + \frac{1}{c_{\nu_a} d^2}
\end{aligned}$$

Bounding underestimation of the optimal arm in D_T: This term contributes only a constant amount to the regret up to time T. To bound it, we invoke Lemma 18 from [12], which we restate below for completeness.

Lemma A.2. *For $T > K$,*

$$\mathbb{E}\left(D_N\right) \leq \begin{cases} (1+\tilde{\eta}) \left(\frac{1}{(\log K)^2} + \frac{\pi^2}{6(\log(1+\tilde{\eta}))^2} \right), \text{ for } \tilde{\eta} > 0 \\ \frac{1+\log(K+1)}{(\log(K+1))^2}, \text{ for } \tilde{\eta} = 0. \end{cases}$$

Combining everything for $\widetilde{\eta} = 0$, we get

$$\mathbb{E}\left(N_a(T)\right) \leq \frac{\log(T)}{\mathrm{KL}_{\inf}\left(\nu_a, \mu_1\right) - d} + O(\log\log(T)) + \frac{1}{c_{\nu_a} d^2} + \frac{1 + \log(K+1)}{(\log(K+1))^2} + 2 \tag{7}$$

The above bound in equation 7 can be optimal over d. Setting d to $(c'_{\nu_a}(\mathrm{KL}_{\inf}(\nu_a, \mu_1))^2 / \log T)^{1/3}$, where $c'_{\nu_a} = 2o(1)/c_{\nu_a}$ we get that

$$\mathbb{E}[N_a(T)] \leq \frac{\log T}{\mathrm{KL}_{\inf}(\nu_a, \mu_1)} + \frac{3(\log T)^{2/3}(c'_{\mu})^{1/3}}{2(\mathrm{KL}_{\inf}(\mu_a, m))^{4/3}} + O((\log T)^{1/3}) + O(\log\log(T))$$

Finally, taking the limit, we get,

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log(T)} \leq \frac{1}{\mathrm{KL}_{\inf}(\nu_a, \mu_1)}.$$

A.3 Proof of Theorem IV.1

Two-arm setting: We first establish the lower bound for the two-armed bandit setting and subsequently extend the argument to the general multi-armed case. Without loss of generality, assume that the mean rewards satisfy $\mu_1 > \mu_2$. Let $\mathcal{M}$ denote an arbitrary class of reward distributions.

Lemma A.3. *Let π be $\mathcal{M}^2$-optimal algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{M}^2$, such that $\gamma \in (0,1)$, and for the second arm,*

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(N_2(T) > (1-\gamma)T)}{\log T} \geq \inf_{\widetilde{\nu}_1 : \mathbb{E}[\widetilde{\nu}_1] \leq \mu_2} - \frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)}$$

Proof of Lemma A.3. Consider a two-armed stochastic bandit problem with an environment $\boldsymbol{\nu} = (\nu_1, \nu_2)$. Introduce an alternative environment $\widetilde{\boldsymbol{\nu}} = (\widetilde{\nu}_1, \nu_2) \in \mathcal{M}^2$ whose arm means satisfy $\widetilde{\mu}_1 < \mu_2$; hence, arm 1 becomes sub-optimal under $\widetilde{\boldsymbol{\nu}}$. For a fixed $\gamma \in (0,1)$, define the event $E = \{N_2(T) > (1-\gamma)T\}$. Applying a change-of-measure argument from $\boldsymbol{\nu}$ to $\widetilde{\boldsymbol{\nu}}$, we obtain

$$\mathbb{P}^{\pi}_{\boldsymbol{\nu}}(E) = \int_E \prod_{t=1}^{T} \frac{dP_{\boldsymbol{\nu}}}{dP_{\widetilde{\boldsymbol{\nu}}}} dP_{\widetilde{\boldsymbol{\nu}}} = \int_E e^{L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})} dP_{\widetilde{\boldsymbol{\nu}}} \tag{8}$$

Here, we define the log-likelihood-ratio process $L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})$ as follows

$$L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}) \;:=\; \log\left(\prod_{t=1}^{T} \frac{dP_{\boldsymbol{\nu}}}{dP_{\widetilde{\boldsymbol{\nu}}}}\big(X_1(t)\big)\right) \;=\; \log\left(\prod_{t=1}^{N_1(T)} \frac{dP_{\nu_1}}{dP_{\widetilde{\nu}_1}}\big(X_1(t)\big)\right) = \sum_{t=1}^{N_1(T)} \log\left(\frac{dP_{\nu_1}}{dP_{\widetilde{\nu}_1}}\big(X_1(t)\big)\right)$$

Now we use the following result from [20], which is stated below for completeness.

Lemma A.4. *Let π be an $\mathcal{M}^K$-optimal algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{M}^K$ and for each sub-optimal arm i,*

$$\frac{N_i(T)}{\log T} \;\xrightarrow[T\to\infty]{P^{\pi}_{\boldsymbol{\nu}}}\; \frac{1}{\mathrm{KL}_{\inf}(\nu_i, \mu_1)}$$

Note that in the above Lemma A.4, the convergence is in terms of probability. Now $L_T(\boldsymbol{\nu}, \boldsymbol{\nu'})$ can be written as follows.

$$L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}) = N_1(T)\, \frac{1}{N_1(T)} \sum_{t=1}^{N_1(T)} \log\left(\frac{dP_{\nu_1}}{dP_{\widetilde{\nu}_1}}\big(X_1(t)\big)\right).$$

Under the environment $\widetilde{\boldsymbol{\nu}}$ as arm 1 is suboptimal thus applying the Lemma A.4 as $T \to \infty$,

$$\frac{N_1(T)}{\log T} \;\xrightarrow[T\to\infty]{P^{\pi}_{\widetilde{\boldsymbol{\nu}}}}\; \frac{1}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)} \tag{9}$$

Under the environment $\widetilde{\boldsymbol{\nu}}$, by the weak law of large numbers as $T \to \infty$,

$$\frac{1}{N_1(T)} \sum_{t=1}^{N_1(T)} \log\left(\frac{dP_{\nu_1}}{dP_{\widetilde{\nu}_1}}\big(X_1(t)\big)\right) \;\xrightarrow{P^{\pi}_{\widetilde{\boldsymbol{\nu}}}}\; -\mathrm{KL}(\widetilde{\nu}_1, \nu_1), \tag{10}$$

Combining equation (9) and (10), we get that under the environment $\widetilde{\boldsymbol{\nu}}$,

$$L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}) = \sum_{t=1}^{N_1(T)} \log\left(\frac{dP_{\nu_1}}{dP_{\widetilde{\nu}_1}}\big(X_1(t)\big)\right) \;\xrightarrow[T\to\infty]{P^{\pi}_{\widetilde{\boldsymbol{\nu}}}}\; -\frac{\log T}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)} \cdot \mathrm{KL}(\widetilde{\nu}_1, \nu_1)$$

Now for any $\varepsilon > 0$ with $\mathbb{P}^{\pi}_{\widetilde{\nu}}$ converging to 1 as $T \to \infty$, the following holds.

$$L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}) \;=\; \sum_{t=1}^{N_1(T)} \log\left(\frac{dP_{\nu_1}}{dP_{\widetilde{\nu}_1}}\big(X_1(t)\big)\right) \;\geq\; -(1+\varepsilon)\frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)} \log T \tag{11}$$

Now, using similar argument as in equation (8), for any $c_T \in \Re$, we can write

$$\begin{aligned}
\mathbb{P}^{\pi}_{\widetilde{\boldsymbol{\nu}}}(E) &= \int_{E \cap \{L_T < c_T\}} e^{-L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})} dP_{\boldsymbol{\nu}} + \int_{E \cap \{L_T \geq c_T\}} e^{-L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})} dP_{\boldsymbol{\nu}} \\
&\leq \int_{\{L_T < c_T\}} e^{-L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})} dP_{\boldsymbol{\nu}} + \int_E e^{-c_T} dP_{\boldsymbol{\nu}} \\
&\leq \mathbb{P}^{\pi}_{\widetilde{\boldsymbol{\nu}}}(\{L_T < c_T\}) + e^{-c_T} \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(E) \\
&\implies \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(E) \geq e^{c_T} \big(\mathbb{P}^{\pi}_{\widetilde{\boldsymbol{\nu}}}(E) - \mathbb{P}^{\pi}_{\widetilde{\boldsymbol{\nu}}}(L_T < c_T)\big)
\end{aligned} \tag{12}$$

Now taking $c_T = -(1+\varepsilon)\frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)} \log T$, then under $\widetilde{\boldsymbol{\nu}}$ we have $\mathbb{P}^{\pi}_{\widetilde{\boldsymbol{\nu}}}(L_T < c_T) \to 0$ and $\mathbb{P}^{\pi}_{\widetilde{\boldsymbol{\nu}}}(E) \to 1$. Now taking log in equation 12, taking $\varepsilon \downarrow 0$ and finally optimizing over the free variable $\widetilde{\nu}_1$ we have,

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(N_2(T) > (1-\gamma)T)}{\log T} \;\geq\; \inf_{\widetilde{\nu}_1 : \mathbb{E}[\widetilde{\nu}_1] \leq \mu_2} -\frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)}$$

□

Multi-arm setting: Now we extend the above result to the multi-armed bandit setting with more than two arms. Without loss of generality, suppose that the means in the environment $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_K)$ satisfy $\mu_1 > \mu_2 > \cdots > \mu_K$. Consider a new environment $\widetilde{\boldsymbol{\nu}} = (\widetilde{\nu}_1, \widetilde{\nu}_2, \ldots, \widetilde{\nu}_{i-1}, \nu_i, \ldots, \nu_K) \in \mathcal{M}^K$ with respective means satisfying $\widetilde{\mu}_j < \mu_i$ for all $j < i$. Under this environment, arm i becomes the optimal arm. Let the event $E = \{N_i(T) > (1-\gamma)T\}$ for some $\gamma \in (0,1)$. Thus, by a change of measure from $\boldsymbol{\nu}$ to $\widetilde{\boldsymbol{\nu}}$, we have

$$\mathbb{P}^{\pi}_{\boldsymbol{\nu}}(E) = \int_E \prod_{t=1}^{T} \frac{dP_{\boldsymbol{\nu}}}{dP_{\widetilde{\boldsymbol{\nu}}}}\big(X_1(t)\big) dP_{\widetilde{\boldsymbol{\nu}}} = \int_E e^{L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})} dP_{\widetilde{\boldsymbol{\nu}}}$$

Similarly, the log-likelihood-ratio process $L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})$ is as follows

$$L_T(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}) \;:=\; \log\Bigg(\prod_{t=1}^{T} \frac{dP_{\boldsymbol{\nu}}}{dP_{\widetilde{\boldsymbol{\nu}}}}\big(X_1(t)\big)\Bigg) \;=\; \log\Bigg(\prod_{j=i}^{i-1} \prod_{t=1}^{N_j(T)} \frac{dP_{\nu_j}}{dP_{\widetilde{\nu}_j}}\big(X_1(t)\big)\Bigg) = \sum_{j=1}^{i-1} L_j(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})$$

For each arm $j \in \{1, 2, \ldots, i-1\}$, define $L_j(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}})$ as follows

$$L_j(\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}) = \sum_{t=1}^{N_j(T)} \log\!\Bigg(\frac{dP_{\nu_j}}{dP_{\widetilde{\nu}_j}}\big(X_1(t)\big)\Bigg)$$

Proceeding along the same lines as in the two-armed bandit analysis, the preceding arguments can be extended to the general multi-armed setting, leading to the following lemma.

Lemma A.5. *Let π be $\mathcal{M}^K$-optimal algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{M}^K$, such that $\gamma \in (0,1)$, for any suboptimal arm-i, the following holds*

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}^{\pi}_{\nu}(N_i(T) > (1-\gamma)T)}{\log T} \;\geq\; -\sum_{j=1}^{i-1} \inf_{\widetilde{\nu}_j : \mathbb{E}[\widetilde{\nu}_j] \leq \mu_i} \frac{\mathrm{KL}(\widetilde{\nu}_j, \nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_j, \mu_i)} \tag{13}$$

To extend the above lemma to the deviation regime $\mathcal{D}_{\gamma}(T) = \big[\log^{1+\gamma}(T),\, (1-\gamma)T\big]$ for any $\gamma \in (0,1)$, we invoke the following result from [20], which is stated below for completeness.

Lemma A.6. *Let $\boldsymbol{\nu}$ be any bandit environment with $B_{\gamma}(T) = [g(T), (1-\gamma)T]$ and any strictly increasing function $g : (1,\infty) \to (0,\infty)$ such that $\lim_{t\to\infty} \frac{g(t)}{\log t} = \infty$ and $g(t) = o(t)$, and let i be a sub-optimal arm in $\boldsymbol{\nu}$.*
Now, if the following condition holds

$$\liminf_{T\to\infty} \frac{\log P^{\pi}_{\boldsymbol{\nu}}\big(N_i(T) > (1-\gamma)T\big)}{\log T} \geq -c_i(\boldsymbol{\nu}).$$

Then,

$$\liminf_{T\to\infty} \inf_{x \in B_{\gamma}(T)} \frac{\log P^{\pi}_{\boldsymbol{\nu}}\big(N_i(T) > x\big)}{\log x} \geq -c_i(\boldsymbol{\nu}).$$

Applying the above lemma in equation (13), we show the desired result as follows:

$$\liminf_{T\to\infty} \inf_{x \in D_{\gamma}(T)} \frac{\log \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(N_i(T) > x)}{\log(x)} \geq \sum_{j=1}^{i-1} \inf_{\widetilde{\nu}_j : \mathbb{E}[\widetilde{\nu}_j] \leq \mu_i} -\frac{\mathrm{KL}(\widetilde{\nu}_j, \nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_j, \mu_i)}$$

A.4 Proof of Theorem IV.5

Two-arm setting: To establish this theorem, we first prove the following lemma in a simplified two-armed bandit setting where arm 1 is optimal, i.e., $\mu_1 > \mu_2$. As introduced earlier, let $\mathcal{L}$ denote the class of reward distributions satisfying Assumptions II.2 and II.3.

Lemma A.7. *Let π be $\mathcal{L}^K$-optimal Generalizeded $\mathrm{KL}_{\inf}$-UCB algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{L}^K$, for $x_T = \log^{1+\gamma}(T)$ such that $\gamma \in (0,1)$, and for the second arm,*

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}^{\pi}_{\boldsymbol{\nu}}(N_2(T) > x_T)}{\log x_T} \;\leq\; -1 \tag{14}$$

Proof of Lemma A.7. Consider a two-armed multi-armed bandit problem with environment $\boldsymbol{\nu} = (\nu_1, \nu_2)$. Without loss of generality, assume $\mu_1 > \mu_2$. Let $\tau_2(m)$ denote the time at which arm 2 is pulled for the m^{th} time, and fix any $\delta \in (0,\, \mu_1 - \mu_2)$. Further, let $C_{\nu} \geq 1$ be a constant satisfying $x_T^{C_{\nu}} < T$. We now derive an upper bound on the following event.

$$
\begin{aligned}
\mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(N_2(T) > x_T\right) &\leq \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (\tau_2(x_T), T] \text{ s.t. } U_1(N_1(t-1), t-1) \leq U_2(N_2(t-1), t-1)\right) \\
&\leq \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (x_T, T] \text{ s.t. } U_1(N_1(t-1), t-1) \leq U_2(x_T, T)\right) \\
&\leq \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (x_T, T] \text{ s.t. } U_1(N_1(t-1), t-1) \leq \mu_2 + \delta\right) \qquad (15) \\
&\quad + \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(U_2(x_T, T) > \mu_2 + \delta\right) \qquad (16) \\
&\leq \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (x_T, T] \text{ s.t. } U_1(N_1(t-1), t-1) \leq \mu_1\right) \\
&\quad + \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(U_2(x_T, T) > \mu_2 + \delta\right) \\
&= \underbrace{\mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } U_1(N_1(t-1), t-1) \leq \mu_1\right)}_{\text{A}} \qquad (17) \\
&\quad + \underbrace{\mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (x_T^{C_\nu}, T] \text{ s.t. } U_1(N_1(t-1), t-1) \leq \mu_1\right)}_{\text{B}} \\
&\quad + \underbrace{\mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(U_2(x_T, T) > \mu_2 + \delta\right)}_{\text{C}}
\end{aligned}
$$

Controlling the Term A and B: To upper bound the term $\boldsymbol{A}$, we invoke Assumption II.2 for the distribution class $\mathcal{L}$ with the exploration function $f_a(t) = \log(t) + 2\log\log(t) + 2\log(1 + N_a(t)) + 1$. The bound then follows through the steps outlined below.

$$
\begin{aligned}
\mathbf{A} &= \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } U_1(N_1(t-1), t-1) \leq \mu_1\right) \\
&= \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\left(\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } \mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) \geq \frac{f_1(t-1)}{N_1(t-1)}\right) \\
&= \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\{\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } N_1(t-1)\mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) - 2\log(1 + N_1(t-1)) - 1 \\
&\qquad\qquad \geq \log(t-1) + \log(\log(t-1))\} \\
&\leq \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\{\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } N_1(t-1)\mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) - 2\log(1 + N_1(t-1)) - 1 \\
&\qquad\qquad \geq \log(x_T) + \log(\log(x_T))\} \\
&\leq \mathbb{P}^{\pi}_{\boldsymbol{\nu}}\{\exists\, t \in \mathbb{N} \text{ s.t. } N_1(t-1)\mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) - 2\log(1 + N_1(t-1)) - 1 \\
&\qquad\qquad \geq \log(x_T) + \log(\log(x_T))\} \\
&\leq \exp(-\log x_T - \log\log x_T) \\
&= \frac{1}{x_T \log x_T}
\end{aligned}
$$

Similarly, by applying the same arguments as above, the probability of the event appearing in term $\boldsymbol{B}$ can be upper bounded in an analogous manner, as detailed below.

$$
\mathbf{B} \leq 1/(x_T^{C_\nu} \log x_T^{C_\nu})
$$

Finally, summing the contributions from the terms $\boldsymbol{A}$ and $\boldsymbol{B}$ and taking the logarithm, we obtain the following bound.

$$
\begin{aligned}
\log(\boldsymbol{A} + \boldsymbol{B}) &\leq -C_\nu \log x_T - \log\log x_T^{C_\nu} + \log(1 + x_T^{C_\nu - 1} C_\nu) \\
&\implies \frac{\log(\boldsymbol{A} + \boldsymbol{B})}{\log x_T} \leq -C_\nu - \frac{\log(C_\nu \log x_T)}{\log x_T} + \frac{\log(1 + x_T^{C_\nu - 1} C_\nu)}{\log x_T} \\
&\implies \limsup_{T \to \infty} \frac{\log(\boldsymbol{A} + \boldsymbol{B})}{\log x_T} \leq -C_\nu + (C_\nu - 1) = -1
\end{aligned}
$$

Controlling Term C: To upper bound the term $\boldsymbol{C}$ above, we apply Assumption II.3 with $d = \mathrm{KL}_{\inf}(\nu_2,\, \mu_2 + \delta) - \frac{f(T)}{x_T}$. Note

that $\frac{f(T)}{x_T} = \frac{1}{f^\gamma(T)}$ it follows that for all sufficiently large T, we have $d > 0$ and for any constant $c_\nu > 0$.

$$\begin{aligned}
\mathbf{C} &= \mathbb{P}_\nu^\pi \left(U_2(x_T, T) > \mu_2 + \delta\right) \\
&= \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(\mathrm{KL}_{\inf}(\hat{\nu}_2(x_T), \mu_2 + \delta) \leq \frac{f(T)}{x_T}\right) \\
&= \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(\mathrm{KL}_{\inf}(\hat{\nu}_2(x_T), \mu_2 + \delta) \leq \mathrm{KL}_{\inf}(\nu_2, \mu_2 + \delta) - \left(\mathrm{KL}_{\inf}(\nu_2, \mu_2 + \delta) - \frac{f(T)}{x_T}\right)\right) \\
&\leq \exp\left[-x_T c_\nu \left(\mathrm{KL}_{\inf}(\nu_2, \mu_2 + \delta) - \frac{f(T)}{x_T}\right)^2\right]
\end{aligned}$$

It can be easily shown that as $T \to \infty$, $\frac{C}{A+B} \to 0$. Finally, summing all the terms, we get

$$\begin{aligned}
\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\boldsymbol{\nu}}^\pi(N_2(T) > x_T)}{\log x_T} &= \limsup_{T\to\infty} \frac{\log(\boldsymbol{A} + \boldsymbol{B} + \boldsymbol{C})}{\log x_T} \\
&= \limsup_{T\to\infty} \frac{\log(\boldsymbol{A} + \boldsymbol{B})}{\log x_T} + \limsup_{T\to\infty} \frac{\log\left(1 + \frac{\boldsymbol{C}}{\boldsymbol{A}+\boldsymbol{B}}\right)}{\log x_T} \\
&\leq \limsup_{T\to\infty} \frac{\log(\boldsymbol{A} + \boldsymbol{B})}{\log x_T} + \limsup_{T\to\infty} \frac{\frac{\boldsymbol{C}}{\boldsymbol{A}+\boldsymbol{B}}}{\log x_T} \\
&\leq -1
\end{aligned}$$

□

Multi-arm setting: We now extend the above argument to the general multi-armed bandit setting with more than two arms. Without loss of generality, assume that $\mu_1 > \mu_2 > \cdots > \mu_K$ in the environment $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_K) \in \mathcal{L}^K$. For any suboptimal arm $i \geq 3$, choose $\delta \in (0, \mu_{i-1} - \mu_i)$. Then, in direct analogy with equations (15) and (16), we obtain the following decomposition:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\nu}}^\pi \left(N_i(T) > x_T\right) &\leq \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(\exists\, t \in (x_T, T] \text{ s.t. } \max_{1\leq j\leq i-1} U_j(N_j(t-1), t-1) \leq \mu_i + \delta\right) \\
&\quad + \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(U_i(x_T, T) > \mu_i + \delta\right) \\
&\leq \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(\forall 1 \leq j \leq i-1, \exists\, t \in (x_T, T] \text{ s.t. } U_j(N_j(t-1), t-1) \leq \mu_i + \delta\right) \\
&\quad + \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(U_i(x_T, T) > \mu_i + \delta\right) \\
&\overset{(a)}{=} \prod_j^{i-1} \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(\exists\, t \in (x_T, T] \text{ s.t. } U_j(N_j(t-1), t-1) \leq \mu_i + \delta\right) \\
&\quad + \mathbb{P}_{\boldsymbol{\nu}}^\pi \left(U_i(x_T, T) > \mu_i + \delta\right)
\end{aligned} \tag{18}$$

(a) is true due to the independence of arms rewards from different arms. Now each of the terms in (18) can be bounded above by the above arguments. Finally, taking log, the product term in (18) becomes a summation. Now we get our desired result for the sub-optimal arm-i as follows

Lemma A.8. *Let π be $\mathcal{L}^K$-optimal Generalizeded $\mathrm{KL}_{\inf}$-UCB algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{L}^K$, for $x_T = \log^{1+\gamma}(T)$ such that $\gamma \in (0, 1)$, and for the i^{th} sub-optimal arm,*

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\boldsymbol{\nu}}^\pi(N_i(T) > x_T)}{\log x_T} \;\leq\; -(i-1) \tag{19}$$

To extend the above lemma to the deviation family $\mathcal{D}_\gamma(T) = \left[\log^{1+\gamma}(T),\, (1-\gamma)T\right]$ for $\gamma \in (0, 1)$ we invoke the following result from [20], which we restate below for completeness.

Lemma A.9. *Let $\boldsymbol{\nu}$ be any bandit environment with $B_\gamma(T) = [g(T), (1-\gamma)T]$ and any strictly increasing function $g : (1, \infty) \to (0, \infty)$ such that $\lim_{t\to\infty} \frac{g(t)}{\log t} = \infty$ and $g(t) = o(t)$, and let i be a sub-optimal arm in $\boldsymbol{\nu}$.*
Now, if the following condition holds

$$\liminf_{T\to\infty} \frac{\log P_{\boldsymbol{\nu}}^\pi\left(N_i(T) > g(T)\right)}{\log g(T)} \leq -c_i(\boldsymbol{\nu}). \tag{20}$$

Then,

$$\liminf_{T\to\infty} \inf_{x\in B_\gamma(T)} \frac{\log P_{\boldsymbol{\nu}}^{\pi}\big(N_i(T) > x\big)}{\log x} \leq -c_i(\boldsymbol{\nu}). \tag{21}$$

Applying the above lemma in equation (19), we show the desired result as follows:

$$\limsup_{T\to\infty} \inf_{x\in\mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}_{\boldsymbol{\nu}}^{\pi}(N_i(T) > x)}{\log x} \;\leq\; -(i-1) \tag{22}$$

A.5 Proof of Corollary V.3

The proof of this corollary proceeds along the same lines as the preceding argument, with minor modifications. Let π denote the empirical KL-UCB algorithm. As shown by [12, Proposition 4], this algorithm is optimal for the bounded-support distribution model $\mathcal{L}_{a,b}^{K}$ when the exploration function is chosen as $f_a(t) = \log t + \log\log t$. Consequently, following the proof of Lemma A.7, we obtain an analogous decomposition and can define the corresponding term $\boldsymbol{A}$, in parallel to equation (17), as follows.

$$\begin{aligned}
\mathbf{A} &= \mathbb{P}_{\boldsymbol{\nu}}^{\pi}\left(\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } U_1(N_1(t-1), t-1) \leq \mu_1\right) \\
&= \mathbb{P}_{\boldsymbol{\nu}}^{\pi}\left(\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } \mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) \geq \frac{f(t-1)}{N_1(t-1)}\right) \\
&= \mathbb{P}_{\boldsymbol{\nu}}^{\pi}\left(\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } N_1(t-1)\mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) \geq \log(t-1) + \log\log(t-1)\right) \\
&= \mathbb{P}_{\boldsymbol{\nu}}^{\pi}\{\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } N_1(t-1)\mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) - 2\log(1+N_1(t-1)) - 1 \\
&\qquad\qquad \geq \log(t-1) + \log(\log(t-1)) - 2\log(1+N_1(t-1)) - 1\} \\
&\overset{(b)}{\leq} \mathbb{P}_{\boldsymbol{\nu}}^{\pi}\{\exists\, t \in (x_T, x_T^{C_\nu}] \text{ s.t. } N_1(t-1)\mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) - 2\log(1+N_1(t-1)) - 1 \\
&\qquad\qquad \geq \log(x_T) + \log(\log(x_T)) - 2\log(1+T-x_T) - 1\} \\
&\leq \mathbb{P}_{\boldsymbol{\nu}}^{\pi}\{\exists\, t \in \mathbb{N} \text{ s.t. } N_1(t-1)\mathrm{KL}_{\inf}(\hat{\nu}_1(t-1), \mu_1) - 2\log(1+N_1(t-1)) - 1 \\
&\qquad\qquad \geq \log(x_T) + \log(\log(x_T)) - 2\log(1+T-x_T) - 1\} \\
&\leq \exp(-\log x_T - \log\log x_T + 2\log(1+T-x_T) + 1) \\
&= \frac{e}{x_T \log x_T (1+T-x_T)^2}
\end{aligned}$$

(b) is true because $N_2(T) > x_T \implies N_1(t-1) \leq N_1(T) \leq T - x_T$. Similarly, following the above steps for the probability of event in term $\boldsymbol{B}$ can also be upper bounded as follows

$$\mathbf{B} \leq \frac{e}{x_T^{C_\nu} \log x_T^{C_\nu} (1+T-x_T^{C_\nu})^2}$$

As $C_\nu \geq 1$, we know $T - x_T^{C_\nu} \leq T - x_T$. Finally summing both the terms in $\boldsymbol{A}$ and $\boldsymbol{B}$ and taking the log, we get the following bound.

$$\begin{aligned}
\log(\boldsymbol{A}+\boldsymbol{B}) &\leq 2 - C_\nu \log x_T - \log\log x_T^{C_\nu} + \log\left(1 + x_T^{C_\nu - 1} C_\nu \left(\frac{1+T-x_T^{C_\nu}}{1+T-x_T}\right)^2\right) \\
&\implies \frac{\log(\boldsymbol{A}+\boldsymbol{B})}{\log x_T} \leq \frac{2}{\log x_T} - C_\nu - \frac{\log(C_\nu \log x_T)}{\log x_T} + \frac{\log(1 + x_T^{C_\nu - 1} C_\nu)}{\log x_T} \\
&\implies \limsup_{T\to\infty} \frac{\log(\boldsymbol{A}+\boldsymbol{B})}{\log x_T} \leq -C_\nu + (C_\nu - 1) = -1
\end{aligned}$$

Proceeding with arguments analogous to those used in the two-armed setting in the proof of Lemma A.7, and subsequently extending them to the multi-armed case, and finally using Lemma A.9 for the deviation family $\mathcal{D}_\gamma(T) = \big[\log^{1+\gamma}(T),\, (1-\gamma)T\big]$, we obtain a similar upper-bound as in Theorem IV.5. Now under discrimination equivalence using the optimal regret tail lower bound from equation (3) we finally get the tight characterization as follows.

For i^{th} sub-optimal arm,

$$\lim_{T\to\infty} \inf_{x\in\mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}_{\boldsymbol{\nu}}^{\pi}(N_i(T) > x)}{\log x} \;=\; -(i-1)$$

A.6 Proof of Theorem VI.1

Two-arm setting: As before, we begin by establishing a simplified version of the upper bound in the two-armed bandit setting, which will subsequently be extended to the general multi-armed case. Without loss of generality, we assume for the remainder of this argument that $\mu_1 > \mu_2$.

Lemma A.10. *Let π be $\mathcal{L}^2_{B,s}$-optimal empirical KL-UCB algorithm. Then, for any environment $\boldsymbol{\nu} \in \mathcal{L}^2_{B,s}$, for $x_T = \log^{1+\gamma}(T)$ such that $\gamma \in (0,1)$, and for the second arm,*

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}^\pi_{\boldsymbol{\nu}}(N_2(T) > x_T)}{\log x_T} \leq \inf_{\widetilde{\nu}_1 : \mathbb{E}[\widetilde{\nu}_1] \leq \mu_2} - \frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)}$$

Proof of Lemma A.10. Consider a 2-arm multi-armed bandit problem with environment $\boldsymbol{\nu} = (\nu_1, \nu_2)$. We consider $\mu_1 > \mu_2$ without the loss of generality. Let's take $\tau_2(m)$ denote the time when arm 2 is played for the m^{th} time. Now for $\delta \in (0, \mu_1 - \mu_2)$,

$$\begin{aligned}
\mathbb{P}^\pi_\nu(N_2(T) > x_T) &\leq \mathbb{P}^\pi_\nu(\exists\, t \in (\tau_2(x_T), T] \text{ s.t. } U_1(N_1(t-1), t-1) \leq U_2(N_2(t-1), t-1)) \\
&\leq \mathbb{P}^\pi_\nu(\exists\, t \in (x_T, T] \text{ s.t. } U_1(N_1(t-1), x_T) \leq U_2(x_T, T)) \\
&\leq \underbrace{\mathbb{P}^\pi_\nu(\exists\, t \in (x_T, T] \text{ s.t. } U_1(N_1(t-1), x_T) \leq \mu_2 + \delta)}_{\mathrm{A}} \\
&\quad + \underbrace{\mathbb{P}^\pi_\nu(U_2(x_T, T) > \mu_2 + \delta)}_{\mathrm{B}}.
\end{aligned}$$

Controlling the first term A: Now in order to upper bound this term we use Sanov's theorem [23], [24] which is stated below for completeness.

Theorem A.11 (Sanov's Theorem)**.** *Let $\mathcal{P}(\Sigma)$ denotes the class of distributions over an underlying set Σ. For $\mathcal{T} \subset \mathcal{P}(\Sigma)$ be a subset of distribution with $\mathcal{T}^0$ and $\overline{\mathcal{T}}$ denoting the interior and closure of $\mathcal{T}$ respectively. Now $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d random variables from drawn from a distribution $\nu \in \mathcal{P}(\Sigma)$. The sequence of the empirical distributions $(\hat{\nu}_n)_{n\in\mathbb{N}}$ satisfy the large deviation principle with rate function $\mathrm{KL}(.,\nu)$ as follows*

$$-\inf_{\nu'\in\mathcal{T}^0} \mathrm{KL}(\nu', \nu) \leq \liminf_{n\to\infty} \log \mathbb{P}(\hat{\nu}_n \in \mathcal{T}) \leq \limsup_{n\to\infty} \log \mathbb{P}(\hat{\nu}_n \in \mathcal{T}) \leq -\inf_{\nu'\in\overline{\mathcal{T}}} \mathrm{KL}(\nu', \nu)$$

The above theorem represents an asymptotic result. However when Σ is a finite set, we get an exact finite sample result with is given in the following equation

$$\mathbb{P}(\hat{\nu}_n \in \mathcal{T}) \leq (n+1)^{|\Sigma|} \exp\{-n \inf_{\nu'\in\overline{\mathcal{T}}} \mathrm{KL}(\nu', \nu)\} \tag{23}$$

In order to apply (23), let us assume there exists a distribution $P^* \in \mathcal{L}_{B,s}$ such that $\mathrm{KL}_{\inf}(P^*, \mu_2 + \delta) = \frac{f(x_T)}{m}$ with $\mathbb{E}[P^*] \leq \mu_2 + \delta$. Now we construct a neighborhood $\mathcal{V}_{P^*}$ around P^* as follows $\mathcal{V}_{P^*} = \{P \in \mathcal{L}_{B,s} : \mathrm{KL}_{\inf}(P, \mu_2 + \delta) \geq \frac{f(x_T)}{m}\}$. Thus, we can further bound the term $\boldsymbol{A}$ as follows.

$$\begin{aligned}
\mathbf{A} &= \mathbb{P}^\pi_\nu(\exists\, t \in (x_T, T] \text{ s.t. } U_1(N_1(t-1), x_T) \leq \mu_2 + \delta) \\
&\leq \mathbb{P}^\pi_\nu(\exists\, t \in \mathbb{N} \text{ s.t. } U_1(N_1(t), x_T) \leq \mu_2 + \delta) \\
&= \mathbb{P}^\pi_\nu(\exists\, m \in \mathbb{N} \text{ s.t. } U_1(m, x_T) \leq \mu_2 + \delta) \\
&\leq \sum_{m=1}^{\infty} \mathbb{P}^\pi_\nu\left(\mathrm{KL}_{\inf}(\hat{\nu}_1(m), \mu_2 + \delta) \geq \frac{f(x_T)}{m}\right) \\
&= \sum_{m=1}^{\infty} \mathbb{P}^\pi_\nu(\hat{\nu}_1(m) \in \mathcal{V}_{P^*}) \\
&\overset{(c)}{\leq} \sum_{m=1}^{\infty} \exp\left[-m \inf_{\nu'\in\bar{\mathcal{V}}_{P^*}} \mathrm{KL}(\nu', \nu_1) + o(m)\right]
\end{aligned}$$

(c) is true by using using the finite version of Sanov's Theorem for finite support distributions with $o(m) = \log(1+m)^s$. Now let $\nu^* \in \mathcal{V}_{P^*}$ be a minimizer of $\inf_{\nu'\in\bar{\mathcal{V}}_{P^*}} \mathrm{KL}(\nu', \nu_1)$. Let's suppose there exists a distribution $\widetilde{\nu} \in \mathcal{L}_{B,s}$ such that $\mathrm{KL}(\nu^*, \nu_1) \geq \mathrm{KL}(\widetilde{\nu}, \nu_1)$. Now we define s_T and C_ν as follows.

$$s_T = \frac{2f(x_T)}{\mathrm{KL}(\widetilde{\nu}, \nu_1)} C_\nu \quad \text{ such that } \quad C_\nu = \inf_{\widetilde{\nu}_1 : \mathbb{E}[\widetilde{\nu}_1] \leq \mu_2 + \delta} \frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2 + \delta)}$$

Notice that $C_\nu \geq 1$ because by definition $\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2 + \delta) \leq \mathrm{KL}(\widetilde{\nu}_1, \nu_1)$. Also, as $\nu^* \in \mathcal{V}_{P^*}$, by definition of $\mathcal{V}^*$, $\mathrm{KL}_{\inf}(\nu^*, \mu_2 + \delta) \geq f(x_T)/m$. Thus for $m > s_T$,

$$\frac{1}{2}\mathrm{KL}(\widetilde{\nu}, \nu_1) \geq \frac{f(x_T)}{m} C_\nu$$

Now the term $\boldsymbol{A}$ can be upper bounded as follows.

$$
\begin{aligned}
\mathbf{A} &\leq \sum_{m=1}^{\infty} \exp\left[-m\mathrm{KL}(\nu^*, \nu_1) + o(m)\right] \\
&= \sum_{m=1}^{s_T} \exp\left[-m\mathrm{KL}(\nu^*, \nu_1) + o(m)\right] + \sum_{m=s_T+1}^{\infty} \exp\left[-m\mathrm{KL}(\nu^*, \nu_1) + o(m)\right] \\
&= \sum_{m=1}^{s_T} \exp\left[-m\mathrm{KL}_{\inf}(\nu^*, \mu_2+\delta) \cdot \frac{\mathrm{KL}(\nu^*, \nu_1)}{\mathrm{KL}_{\inf}(\nu^*, \mu_2+\delta)} + o(m)\right] \\
&\qquad\qquad + \sum_{m=s_T+1}^{\infty} \exp\left[-m\mathrm{KL}(\nu^*, \nu_1) + o(m)\right] \\
&\leq \sum_{m=1}^{s_T} \exp\left[-f(x_T) \cdot \inf_{\widetilde{\nu}_1 : \mathbb{E}[\widetilde{\nu}_1] \leq \mu_2+\delta} \frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2+\delta)} + o(m)\right] \\
&\qquad\qquad + \sum_{m=s_T+1}^{\infty} \exp\left[-m\mathrm{KL}(\nu^*, \nu_1) + o(m)\right] \\
&\leq \sum_{m=1}^{s_T} \exp\left[-f(x_T) \cdot C_\nu + o(m)\right] + \sum_{m=s_T+1}^{\infty} \exp\left[-m\mathrm{KL}(\widetilde{\nu}, \nu_1) + o(m)\right] \\
&\leq \sum_{m=1}^{s_T} \exp\left[-f(x_T) \cdot C_\nu\right] \exp[o(m)] + \sum_{m=s_T+1}^{\infty} \exp\left[-f(x_T) C_\nu - \frac{m}{2} \cdot \mathrm{KL}(\widetilde{\nu}, \nu) + o(m)\right] \\
&= \exp\left[-f(x_T) \cdot C_\nu\right] \left(\sum_{m=1}^{s_T} \exp[o(m)] + \sum_{m=s_T+1}^{\infty} \exp\left[-\frac{m}{2} \cdot \mathrm{KL}(\widetilde{\nu}, \nu) + o(m)\right] \right)
\end{aligned}
$$

Controlling the second term B: The term $\boldsymbol{B}$ can be upper bounded using Assumption II.3 with $d = \mathrm{KL}_{\inf}(\nu_2, \mu_2+\delta) - \frac{f(T)}{x_T}$. Note that $\frac{f(T)}{x_T} = \frac{1}{f^\gamma(T)}$ it follows that for all sufficiently large T, we have $d > 0$ and for any constant $c_\nu > 0$.

$$
\begin{aligned}
\mathbf{B} &= \mathbb{P}_\nu^\pi \left(U_2(x_T, T) > \mu_2 + \delta\right) \\
&= \mathbb{P}_\nu^\pi \left(\mathrm{KL}_{\inf}(\hat{\nu}_2(x_T), \mu_2+\delta) \leq \frac{f(T)}{x_T}\right) \\
&= \mathbb{P}_\nu^\pi \left(\mathrm{KL}_{\inf}(\hat{\nu}_2(x_T), \mu_2+\delta) \leq \mathrm{KL}_{\inf}(\nu_2, \mu_2+\delta) - \left(\mathrm{KL}_{\inf}(\nu_2, \mu_2+\delta) - \frac{f(T)}{x_T}\right)\right) \\
&\leq \exp\left[-x_T c_\nu \left(\mathrm{KL}_{\inf}(\nu_2, \mu_2+\delta) - \frac{f(T)}{x_T}\right)^2\right]
\end{aligned}
$$

Finally, summing the contributions from the terms $\boldsymbol{A}$ and $\boldsymbol{B}$ and taking the logarithm, we obtain the following bound.

$$
\begin{aligned}
\log(\mathbf{A}+\mathbf{B}) \leq -f(x_T) \cdot C_\nu &+ \log\Bigg(\underbrace{\sum_{m=1}^{s_T} \exp[o(m)] + \sum_{m=s_T+1}^{\infty} \exp\left[-\tfrac{m}{2} \cdot \mathrm{KL}(\widetilde{\nu}, \nu) + o(m)\right]}_{} \\
&\underbrace{+ \exp\left[-x_T c_\nu \left(\mathrm{KL}_{\inf}(\nu_2, \mu_2+\delta) - \tfrac{f(T)}{x_T}\right)^2 + f(x_T) C_\nu\right]}_{\mathrm{G}} \Bigg)
\end{aligned}
$$

For s-finite support distribution $o(m) = \log(m+1)^s$ we have,

$$
\begin{aligned}
\frac{\log G}{f(x_T)} &= \frac{1}{f(x_T)} \log \Bigg(\sum_{m=1}^{s_T} \exp[o(m)] + \sum_{m=s_T+1}^{\infty} \exp\left[-\tfrac{m}{2} \cdot \mathrm{KL}(\widetilde{\nu}, \nu) + o(m)\right] \\
&\qquad + \exp\left[-x_T \cdot c_\nu \left(\mathrm{KL}_{\inf}(\nu_2, \mu_1) - \frac{f(T)}{x_T} \right)^2 + C_\nu f(x_T) \right] \Bigg) \\
&= \frac{1}{f(x_T)} \log \Bigg(\sum_{m=1}^{s_T} (m+1)^s + \sum_{m=s_T+1}^{\infty} \exp\left[-\tfrac{m}{2} \cdot \mathrm{KL}(\widetilde{\nu}, \nu) + s \log(m+1)\right] \\
&\qquad + \exp\left[-x_T \cdot c_\nu \left(\mathrm{KL}_{\inf}(\nu_2, \mu_1) - \frac{f(T)}{x_T} \right)^2 + C_\nu f(x_T) \right] \Bigg) \\
&\leq \frac{1}{f(x_T)} \log \Bigg(s_T (s_T+1)^s + O(\exp(-(s_T+1))) \\
&\qquad + \exp\left[-x_T \cdot c_\nu \left(\mathrm{KL}_{\inf}(\nu_2, \mu_1) - \frac{f(T)}{x_T} \right)^2 + C_\nu f(x_T) \right] \Bigg) \\
&= \frac{1}{f(x_T)} \log \Bigg(O(f^{s+1}(x_T)) + O(\exp(-f(x_T))) \\
&\qquad + \exp\left(-x_T \cdot c_\nu \mathrm{KL}^2_{\inf}(\nu_2, \mu_1) + 2 c_\nu \mathrm{KL}_{\inf}(\nu_2, \mu_1) f(T) - c_\nu f^{1-\gamma}(T) + C_\nu f(x_T) \right) \Bigg) \qquad (24)
\end{aligned}
$$

Note that as $T \to \infty$ the second and third term inside the log in equation (24) goes to zero and finally we get

$$
\limsup_{T \to \infty} \frac{\log G}{f(x_T)} \to \frac{\log O(f(x_T)^{s+1})}{f(x_T)} \to 0
$$

Finally taking $\delta \downarrow 0$, we get the desired result as follows

$$
\limsup_{T \to \infty} \frac{\log \mathbb{P}^\pi_{\boldsymbol{\nu}}(N_2(T) > x_T)}{\log x_T} \;\leq\; \inf_{\widetilde{\nu}_1 : \mathbb{E}[\widetilde{\nu}_1] \leq \mu_2} - \frac{\mathrm{KL}(\widetilde{\nu}_1, \nu_1)}{\mathrm{KL}_{\inf}(\widetilde{\nu}_1, \mu_2)}
$$

□

Multi-arm Setting: The extension to the multi-arm setting follows along lines analogous to the same extension in the proof of Theorem IV.5 in Section A. Finally, invoking Lemma A.9, we obtain the following result for the deviation family $\mathcal{D}_\gamma(T) = \left[\log^{1+\gamma}(T),\, (1-\gamma)T\right]$ with $\gamma \in (0,1)$, as stated below.

$$
\limsup_{T \to \infty} \inf_{x \in \mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}^\pi_{\boldsymbol{\nu}}(N_i(T) > x)}{\log x} \leq - \sum_{j=1}^{i-1} \inf_{\substack{\widetilde{\nu} \in \mathcal{M}: \\ m(\widetilde{\nu}) < \mu_i}} \frac{\mathrm{KL}(\widetilde{\nu}, \nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu}, \mu_i)}
$$

Now using the optimal regret tail lower bound from equation (3) we finally get the tight characterization as follows.

$$
\lim_{T \to \infty} \inf_{x \in \mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}^\pi_{\boldsymbol{\nu}}(N_i(T) > x)}{\log x} = - \sum_{j=1}^{i-1} \inf_{\substack{\widetilde{\nu} \in \mathcal{M}: \\ m(\widetilde{\nu}) < \mu_i}} \frac{\mathrm{KL}(\widetilde{\nu}, \nu_j)}{\mathrm{KL}_{\inf}(\widetilde{\nu}, \mu_i)}
$$

B. Relevant literature

Expected regret minimization in the stochastic multi-armed bandit problem has a long and rich history, dating back to the early work of [26] in the context of clinical trials and [27]. A foundational result was achieved by [1], who established a fundamental lower bound on the expected cumulative regret for parametric reward models. This bound shows that the expected regret must grow at least logarithmically in the time horizon T, with a leading constant determined by the Kullback-Leibler (KL) divergences between the optimal arm's and each sub-optimal arm's reward distributions. This result was later generalized by [2]. Algorithms that achieve this lower bound are often called optimal. Among the many optimal algorithms proposed, one of the prominent algorithms is the KL-upper confidence bound (KL-UCB) algorithm proposed by [9]. Even though the different variants of UCB-type algorithm have long been studied [28]–[30], [9] provided finite time regret analysis of the KL-UCB algorithm and first showed that the family of KL-UCB algorithm is asymptotically optimal not only for single parameter exponential families (SPEFs), but also the empirical KL-UCB is optimal for generic finite and bounded supported reward

distributions and conjectured that this is also optimal for generic bounded rewards. Later, [12] formally proved this conjecture, making the KL-UCB algorithm optimal for generic bounded support distributions. Since then, a predominant focus of the subsequent literature has been the design of optimal algorithms, especially for well-structured distribution families [3]–[12]. We refer the reader to [14] for a comprehensive survey.

Looking beyond expected regret has attracted sustained interest with early works [15], [16] examining deviation and concentration properties of regret distribution. More recently, a growing body of work has undertaken a systematic study of the distributional properties of regret, including worst-case behavior [17]. Consequently, [18] studies some typical behavior and fluctuations of regret when T is large, developing strong laws of large numbers and central limit theorems for bandit algorithms, for instance dependent settings. Further, some atypical behavior of regret is studied by [20], demonstrating that algorithms which are asymptotically optimal in expected regret, namely, those that attain the Lai-Robbins lower bound, may nonetheless exhibit heavy or poorly controlled regret tails. This is one of the fragility issues of optimal bandit algorithms, highlighting a fundamental limitation of expectation-based optimality criteria.