OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

The Fragility of Optimized Bandit Algorithms

Lin Fan
Kellogg School of Management, Northwestern University, Evanston, IL 60208, lin.fan@kellogg.northwestern.edu

Peter W. Glynn
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, glynn@stanford.edu

Much of the literature on optimal design of bandit algorithms is based on minimization of expected regret. It is well known that algorithms that are optimal over certain exponential families can achieve expected regret that grows logarithmically in the number of trials, at a rate specified by the Lai-Robbins lower bound. In this paper, we show that when one uses such optimized algorithms, the resulting regret distribution necessarily has a very heavy tail, specifically, that of a truncated Cauchy distribution. Furthermore, for $p > 1$, the p'th moment of the regret distribution grows much faster than poly-logarithmically, in particular as a power of the total number of trials. We show that optimized UCB algorithms are also fragile in an additional sense, namely when the problem is even slightly mis-specified, the regret can grow much faster than the conventional theory suggests. Our arguments are based on standard change-of-measure ideas, and indicate that the most likely way that regret becomes larger than expected is when the optimal arm returns below-average rewards in the first few arm plays, thereby causing the algorithm to believe that the arm is sub-optimal. To alleviate the fragility issues exposed, we show that UCB algorithms can be modified so as to ensure a desired degree of robustness to mis-specification. In doing so, we also show a sharp trade-off between the amount of UCB exploration and the heaviness of the resulting regret distribution tail.

Key words: Multi-armed Bandits, Regret Distribution, Limit Theorems, Model Mis-specification, Robustness

History: To appear in Operations Research (first version on arXiv: Sept 28, 2021)

1. Introduction

The multi-armed bandit (MAB) problem is a widely studied model that is both useful in practical applications and is a valuable theoretical paradigm exhibiting the exploration-exploitation trade-off that arises in sequential decision-making under uncertainty. More specifically, the goal in a MAB problem is to maximize the expected reward derived from playing, at each time step, one of K bandit arms. Each arm has its own unknown reward distribution, so that playing a particular arm both provides information about that arm's reward distribution (exploration) and provides an associated random reward (exploitation). One measure of the quality of a MAB algorithm is the (pseudo-)regret $R(T)$, which is essentially the number of times the sub-optimal arms are played

over a time horizon T, as compared to an oracle that acts optimally with knowledge of the means of all arm reward distributions; a precise definition will be given in Section 2.

There is an enormous literature on this problem, with much of the research having been focused on algorithms that attempt to minimize expected regret. In this regard, a fundamental result is the Lai-Robbins lower bound that establishes that the expected regret $\mathbb{E}[R(T)]$ grows logarithmically in T, with a multiplier that depends on the Kullback-Leibler (KL) divergences between the optimal arm and each of the sub-optimal arms; see Lai and Robbins (1985). A predominant focus in the bandit literature is on designing algorithms that attain the Lai-Robbins lower bound over particular exponential families of distributions; see Lai and Robbins (1985) and Burnetas and Katehakis (1996). We call such algorithms *optimized*. Among the many optimized algorithms in the literature, two prominent examples are the KL-upper confidence bound (KL-UCB) algorithm and Thompson sampling (TS); see Cappé et al. (2013) (and earlier work: Garivier and Cappé (2011), Maillard et al. (2011)) for KL-UCB, and Korda et al. (2013) for TS (originally proposed by Thompson (1933)). Earlier optimized UCB-type algorithms can be found in, for example, Lai (1987) and Agrawal (1995).

In this paper, we show that any optimized algorithm necessarily has the undesirable property that the tail of $R(T)$ is very heavy. In particular, because $\mathbb{E}[R(T)]$ is $O(\log(T))$ (where $O(a_T)$ is any sequence having the property that its absolute value is dominated by a constant multiple of a_T), Markov's inequality implies that for $c>0$, $\mathbb{P}(R(T)>cT)=O(\log(T)/T)$ as $T\to\infty$. One of our central results is a lower bound characterization of $\mathbb{P}(R(T)>cT)$ that roughly establishes that this probability is attained, namely it is roughly of order T^{-1} for optimized algorithms. More precisely, our Theorem 1 shows that optimized MAB algorithms automatically have the property that

$$\mathbb{P}(R(T)>x)\asymp\frac{1}{x}$$

as $T\to\infty$, uniformly in x with $T^a\le x\le cT$, for any $0<a<1$ and suitable $c>0$. (We write $a_T\asymp b_T$ as $T\to\infty$ whenever $\log(a_T)/\log(b_T)$ converges to 1 as $T\to\infty$.) In other words, the tail of the regret $R(T)$ looks, in logarithmic scale, like that of a *truncated Cauchy distribution* (truncated due to the time horizon T). Thus, such algorithms fail to produce logarithmic regret with large probability, and when they fail to produce such regret, the magnitude of the regret can be very large. This is one sense in which bandit algorithms optimized for expected regret can be fragile.

An additional sense in which such optimized bandit algorithms are fragile is their sensitivity to model mis-specification. By this, we mean that if an algorithm has been optimized to attain the Lai-Robbins lower bound over a particular class of bandit environments (e.g., with the arm distributions belonging to a specific exponential family), then we can see much worse regret behavior when the

environment presented to the algorithm does not belong to the class. For example, we show that for the KL-UCB algorithm designed for Gaussian environments with known and equal variances but unknown means, the expected regret for Gaussian environments can grow as a power T^r when the variance of the optimal arm's rewards is larger than the variance built into the algorithm's design. In fact, r can be made arbitrarily close to 1 depending on how large the optimal arm's variance is, relative to the variance of the algorithm's design (Corollary 2). In other words, even when the mis-specification remains Gaussian, the expected regret can grow at a rate close to linear in the time horizon T. Besides mis-specification of the bandits' marginal reward distributions, optimized algorithms are equally susceptible to mis-specification of the serial dependence structure of rewards. For example, expected regret deteriorates similarly as reward processes (e.g., evolving as Markov chains) become more autocorrelated (Corollary 3, Proposition 5 and Example 5).

A final sense in which such optimized algorithms are fragile is that when one only slightly modifies the objective, the regret behavior of the algorithm can look much worse. In particular, suppose that we consider minimizing $\mathbb{E}[R(T)^p]$ for some $p>1$, rather than $\mathbb{E}[R(T)]$. This objective would arise naturally, for example, in the presence of risk aversion to high regret. One might reasonably expect that algorithms optimized for $\mathbb{E}[R(T)]$ would have the property that $\mathbb{E}[R(T)^p]$ would then grow poly-logarithmically in T. However, the Cauchy-type tails discussed earlier imply that $(R(T)/\log(T))^p$ is not a uniformly integrable sequence. We show in Corollary 4 that for optimized algorithms, $\mathbb{E}[R(T)^p]$ grows roughly at least as fast as T^{p-1} as $T\to\infty$.

Our proofs rely on change-of-measure arguments that also provide insight into how algorithms optimized for expected regret can fail to identify the optimal arm, thereby generating large regret. For example, we show that conditional on large regret, the sample means of sub-optimal arms obey laws of large numbers that indicate that they continue to behave in their usual way; see Proposition 3. This suggests that the most likely way that large regret occurs for such optimized algorithms is when the optimal arm under-performs in the exploration phase at the start of the experiment, after which it is played infrequently, thereby generating large amounts of regret. This intuitive scenario has been heuristically considered several times in the literature (see, e.g., Audibert et al. (2009)), but this paper provides the theoretical justification for its central role in generating large regret.

To mitigate some of the fragility issues we expose, we show how to modify UCB algorithms so that their regret tails are lighter. By suitably increasing the rate of UCB exploration, we can achieve any polynomial or exponential rate of tail decay; see Proposition 7. For example, in well-specified settings, if one increases the nominal amount of exploration by a factor of $1+b$ times for any desired $b>0$, then the tail of the resulting regret distribution will have an exponent of $-(1+b)$ (or less). In particular, $\mathbb{P}(R(T)>x)\asymp x^{-(1+b)}$ as $T\to\infty$, uniformly in x with $\log^a(T)\le x\le cT$, for any $a>1$ and suitable $c>0$. By lightening the regret distribution tail to a given exponent, we also

create a prescribed margin of safety against model mis-specification. The modified UCB algorithm becomes more robust to mis-specification of the reward distribution (Corollary 5) and of the serial dependence structure of rewards (Corollary 6). Of course, these benefits must come at the cost of greater expected regret. We study (sharp) trade-offs between the regret tail and expected regret in Proposition 7 and Theorem 4.

Our study of the tail of the regret distribution of MAB algorithms, and our uncovering of the above fragility phenomena, underscore the value of understanding the regret distribution beyond the expected regret performance measure that the literature focuses overwhelmingly on. Despite the fundamental role of expected regret for sequential decision-making under uncertainty, as we show, important insights can be missed and severe fragility can result from optimizing for expected regret alone. Our work thus provides a novel and useful complement to the by-now mature theory of expected regret in the bandit literature. See also Section 1.1 for recent related work concerning the regret distribution.

The rest of the paper is structured as follows. After discussing related work in Section 1.1, we introduce the setup for the paper in Section 2. In Section 3.1, we establish our main result, Theorem 1, that optimized algorithms have regret distributions for which the tails are truncated Cauchy. This result requires a technical condition (Definition 3), which holds essentially for all continuous reward distributions. To illustrate the key ideas behind Theorem 1, we prove a simplified version of the result in Section 3.2. We develop in Section 3.3 tight upper bounds characterizing the regret tail for KL-UCB in settings where the regret tail is lighter than truncated Cauchy (because the condition in Definition 3 does not hold); see Theorem 2. In Section 4.1, we discuss the connections between the heavy regret tails of optimized algorithms and their susceptibility to model mis-specification. Afterwards, we show in Sections 4.2 and 4.4 that the performance of optimized algorithms can deteriorate sharply under the slightest amount of mis-specification of the distribution or the serial dependence structure of the rewards. These insights make use of results from Section 4.3, where we establish general lower bounds for the regret tail of algorithms such as KL-UCB when the rewards come from stochastic processes; see Theorem 3. Moreover, we show in Section 4.5 that such optimized algorithms offer no control over the p'th moment of regret for any $p > 1$. We extend the regret tail characterizations for exponential family models in Theorem 1 to models with general reward distributions in Section 5.1. In Section 5.2, we develop a trade-off in Theorem 4 showing that lighter regret tails come at the cost of greater expected regret. Our result significantly generalizes the Lai-Robbins lower bound for expected regret (as well as the Burnetas-Katehakis extension). In Section 6.1, building upon Section 3.3, we discuss how to modify UCB algorithms to achieve any desired regret tail, with polynomial or exponential rates of decay, by suitably increasing the rate of exploration. We then discuss how the modifications provide protection against mis-specification

of the distribution of rewards and of the serial dependence structure of rewards in Sections 6.2 and 6.3, respectively. In Section 7, we examine some numerical experiments. We conclude with the proofs of Theorems 1, 2, 3 and 4 in Appendices A, B, C and D, respectively.

1.1. Related Work

In terms of related work, Audibert et al. (2009), Salomon and Audibert (2011) study concentration properties of the regret distribution. In particular, Audibert et al. (2009) develop a finite-time upper bound on the tail of the regret distribution for a particular version of UCB in bounded reward settings. Their upper bound has polynomial rates of tail decay, which are adjustable depending on algorithm settings. One of their motivations for developing regret tail bounds is to establish a trade-off between the rate of exploration and the resulting heaviness of the regret tail. However, it is lower bounds on the regret tail that are needed to conclusively establish the trade-off and confirm that the regret distribution is heavy-tailed. Our lower bounds turn out to be frequently tight.

The regret distribution tail approximations developed in the current work are complementary to the strong laws of large numbers (SLLN's) and central limit theorems (CLT's) developed for bandit algorithms in instance-dependent settings in Fan and Glynn (2022). For example, in the Gaussian bandit setting (with unit variances for simplicity), for both TS and UCB, the regret satisfies the SLLN:

$$\frac{R(T)}{\log(T)} \xrightarrow{\text{a.s.}} \sum_{i \neq i^*} \frac{2}{\Delta_i}$$

and the CLT:

$$\frac{R(T) - \sum_{i \neq i^*} \frac{2}{\Delta_i} \log(T)}{\sqrt{\sum_{i \neq i^*} \frac{8}{\Delta_i^2} \log(T)}} \Rightarrow N(0,1),$$

where $\Delta_i > 0$ is the difference between the mean of the optimal arm i^* and that of sub-optimal arm i, and $\Rightarrow$ denotes convergence in distribution. These results can be viewed as describing the typical behavior and fluctuation of regret when T is large. This stands in contrast to the results in the current work, which describe the tail behavior of the regret. Tails are generally affected by atypical behavior. As noted above, our arguments show that the regret tail is impacted by trajectories on which the algorithm mis-identifies the optimal arm. The mean and the variance in the CLT both scale as $\log(T)$ with the time horizon T. By analogy with the large deviations theory for sums of independent, identically distributed (iid) random variables, this suggests that large deviations of regret correspond to deviations from the expected regret that are of order $\log(T)$. We characterize

the tail of the regret beyond $\log^{1+\gamma}(T)$ for small $\gamma > 0$, and we save the analysis of deviations on the $\log(T)$ scale for future work.

The regret distribution of MAB algorithms has also been studied in an asymptotic regime different from the Lai and Robbins regime that this paper and Fan and Glynn (2022) focus on. Kuang and Wager (2021), Fan and Glynn (2021) and Kalvit and Zeevi (2021) obtain diffusion approximations for the regret distribution of MAB algorithms (including for TS and UCB) in the setting where the gap between arm means scales like $1/\sqrt{T}$, with the time horizon $T \to \infty$.

Recently, Ashutosh et al. (2021) showed that for an algorithm to achieve expected regret growing logarithmically in the time horizon across a collection of environment instances, the distributional class of arm rewards cannot be too large. For example, if the rewards are known to be sub-Gaussian, then an upper bound restriction on the variance proxy is required. They conclude that if such a restriction is mis-specified, then the worst case (for some environment instance) expected regret could grow polynomially in the time horizon. Their result provides no information about algorithm behavior for any particular environment instance, nor does it cover narrower classes of distributions (e.g., Gaussian).

Our results on model mis-specification are significantly differentiated from other results on model mis-specification in the bandit literature. We consider the effect of mis-specifying, for example, the distribution or serial dependence structure of rewards, on the (frequentist) regret tail (and as a direct consequence, also the expected regret). From the perspective of Bayesian expected regret, Simchowitz et al. (2021) studies mis-specification of the prior distribution used by algorithms, and Liu et al. (2022) studies mis-specified Bernoulli bandits for algorithms based on Gaussian prior and likelihood structure. From the perspective of frequentist expected regret, existing literature, including Ghosh et al. (2017), Lattimore et al. (2020), Foster et al. (2020), Takemura et al. (2021) and Krishnamurthy et al. (2021), study the effect of mis-specifying a linear regression structure when the true regression function is not exactly linear.

There is also a growing literature on risk-averse formulations of the MAB problem, with a non-comprehensive list being: Sani et al. (2012), Maillard (2013), Galichet et al. (2013), Zimin et al. (2014), Szorenyi et al. (2015), Vakili and Zhao (2016), Cassel et al. (2018), Tamkin et al. (2019), Zhu and Tan (2020), Prashanth et al. (2020), Baudry et al. (2021), Khajonchotpanya et al. (2021). As noted earlier, risk-averse formulations involve defining arm optimality using criteria other than the expected reward. These papers consider mean/variance criteria, value-at-risk, or conditional value-at-risk measures, and develop algorithms which achieve good (or even optimal in some cases) regret performance relative to their chosen criterion. Our results serve as motivation for these papers, and highlight the need to consider robustness in many MAB problem settings. We believe it would be interesting to investigate in follow-up studies the fragility issues exposed in our paper through the lens of these risk-averse MAB formulations.

2. Model and Preliminaries

2.1. The Multi-armed Bandit Framework

A K-armed MAB evolves within a bandit environment $\nu = (Q_1, \ldots, Q_K)$, where each Q_i is a distribution on $\mathbb{R}$. At time t, the decision-maker selects an arm $A(t) \in [K] := \{1, \ldots, K\}$ to play. Upon selecting the arm $A(t)$, a reward $Y(t)$ from arm $A(t)$ is received as feedback. The conditional distribution of $A(t)$ given $A(1), Y(1), \ldots, A(t-1), Y(t-1)$ is $\pi_t(\cdot \mid A(1), Y(1), \ldots, A(t-1), Y(t-1))$, where $\pi = (\pi_t, t \geq 1)$ is a sequence of probability kernels, which constitutes the bandit algorithm. For each $t \geq 1$, π_t is a probability kernel from $([K] \times \mathbb{R})^{t-1}$ to $[K]$. The conditional distribution of $Y(t)$ given $A(1), Y(1), \ldots, A(t-1), Y(t-1), A(t)$ is $Q_{A(t)}(\cdot)$. We write $X_i(s)$ to denote the reward received when arm i is played for instance s, so that $Y(t) = X_{A(t)}(N_{A(t)}(t))$, where $N_i(t) = \sum_{s=1}^{t} \mathbb{I}(A(s) = i)$ denotes the number of plays of arm i up to and including time t.

For any time t, the interaction between the algorithm π and the environment ν induces a unique probability $\mathbb{P}_{\nu\pi}(\cdot)$ on $([K] \times \mathbb{R})^{\infty}$ for which

$$\mathbb{P}_{\nu\pi}(A(1) = a_1, Y(1) \in dy_1, \ldots, A(t) = a_t, Y(t) \in dy_t) = \prod_{s=1}^{t} \pi_s(a_s \mid a_1, y_1, \ldots, a_{s-1}, y_{s-1}) Q_{a_s}(dy_s).$$

Here, dy denotes an infinitesimal set containing y. For $t \geq 1$, we write $\mathbb{E}_{\nu\pi}[\cdot]$ to denote the expectation associated with $\mathbb{P}_{\nu\pi}(\cdot)$.

The quality of an algorithm π operating in an environment $\nu = (Q_1, \ldots, Q_K)$ is measured by the (pseudo-)regret (at time T):

$$R(T) = \sum_{i=1}^{K} N_i(T) \Delta_i,$$

where $\Delta_i = \mu_*(\nu) - \mu(Q_i)$ and $\mu_*(\nu) = \max_{Q \in \nu} \mu(Q)$. (For any distribution Q, we use $\mu(Q)$ to denote its mean.) An arm i is called optimal if $\Delta_i = 0$, and sub-optimal if $\Delta_i > 0$. The goal in most settings is to find an algorithm π which minimizes the expected regret $\mathbb{E}_{\nu\pi}[R(T)]$, i.e., plays the optimal arm(s) as often as possible in expectation.

When discussing the regret distribution tail in multi-armed settings, we will often reference (for any given environment) the i-th-best arm (with the i-th largest mean). For each $i = 1, \ldots, K$, we will use $r(i) \in [K]$ to denote the index/label of the i-th-best arm. To keep our discussions and derivations streamlined, unless specified otherwise, throughout the paper we will only consider environments where for each $i = 1, \ldots, K$, the i-th-best arm is unique.

2.2. Optimized Algorithms

In order to discuss optimized algorithms, we consider arm reward distributions from a one-dimensional exponential family, parameterized by mean, of the form:

$$P^z(dx) = \exp\big(\theta_P(z) \cdot x - \Lambda_P(\theta_P(z))\big) P(dx), \quad z \in \mathcal{I}_P. \tag{1}$$

Here, P is a base distribution with cumulant generating function (CGF) Λ_P. We use $\mathcal{I}_P$ to denote the set of all possible means for distributions P^z of the form in (1), with $\theta_P(z)$ being any real number in the set $\Theta_P = \{\theta \in \mathbb{R} : \Lambda_P(\theta) < \infty\}$. Moreover, for each $z \in \mathcal{I}_P$, we use $\theta_P(z)$ to denote the unique value for which $\mu(P^z) = z$. (Also recall that $\Lambda_P'(\theta_P(z)) = z$.) Throughout the paper, we will always work with base distributions P such that Θ_P contains a neighborhood of zero.

For a base distribution P, we denote the mean-parameterized model in (1) via:

$$\mathcal{M}_P = \{P^z : z \in \mathcal{I}_P\}, \tag{2}$$

which induces a class $\mathcal{M}_P^K$ of K-armed bandit environments, where each environment consists of a K-tuple of distributions from $\mathcal{M}_P$. The KL divergence between distributions in $\mathcal{M}_P$ with means $z_1, z_2 \in \mathcal{I}_P$ is denoted by $d_P(z_1, z_2)$, and can be expressed as:

$$\begin{aligned} d_P(z_1, z_2) &= \int \log \frac{dP^{z_1}}{dP^{z_2}}(x)\, P^{z_1}(dx) \\ &= \Lambda_P(\theta_P(z_2)) - \Lambda_P(\theta_P(z_1)) - \Lambda_P'(\theta_P(z_1)) \cdot \big(\theta_P(z_2) - \theta_P(z_1)\big), \end{aligned} \tag{3}$$

where dP^{z_1}/dP^{z_2} denotes the likelihood ratio of P^{z_1} to P^{z_2}.

From the seminal work of Lai and Robbins (1985), there is a precise characterization of the minimum possible growth rate of expected regret for an algorithm π designed for $\mathcal{M}_P^K$. We start with the notion of *consistency* in Definition 1 below, which restricts the types of algorithms considered in order to formulate a theory of optimality. This notion rules out unnatural algorithms that over-specialize and perform very well in particular environment instances within a class, but very poorly in other instances.

Definition 1 ($\mathcal{M}_P$-Consistent Algorithm). An algorithm π is $\mathcal{M}_P$-consistent if for any $a > 0$, any environment $\nu \in \mathcal{M}_P^K$, and each sub-optimal arm i:

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{T^a} = 0. \tag{4}$$

The Lai-Robbins lower bound is then formulated for consistent algorithms. In particular, for any $\mathcal{M}_P$-consistent algorithm π, any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ and each sub-optimal arm i,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{\log(T)} \geq \frac{1}{d_P(\mu_i, \mu_*(\nu))}. \tag{5}$$

We say that an $\mathcal{M}_P$-consistent algorithm π is $\mathcal{M}_P$-*optimized* if the lower bound in (5) is achieved, as in the following Definition 2.

Definition 2 ($\mathcal{M}_P$-Optimized Algorithm). An algorithm π is $\mathcal{M}_P$-optimized if for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ and each sub-optimal arm i,

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{\log(T)} = \frac{1}{d_P(\mu_i, \mu_*(\nu))}.$$

3. Characterization of the Regret Tail

3.1. Truncated Cauchy Tails

In this section, we show that for many classes of exponential family bandit environments, the tail of the regret distribution of optimized algorithms is essentially that of a truncated Cauchy distribution. Moreover, for such classes, the tail is truncated Cauchy for *every environment* within the class. This is established in Theorem 1. As we will see, this truncated Cauchy tail property always holds when the exponential family is continuous with left tails that are lighter than exponential (possessing CGF's that are finite on the negative half of the real line). When the exponential family is discrete or has exponential left tails, the regret distribution tail is generally lighter than truncated Cauchy, but still heavy and decaying at polynomial rates.

As discussed in the Introduction, the regret tail characterization that we develop here reveals several important insights about the fragility of optimized bandit algorithms. For example, when the regret tail is truncated Cauchy, as is generally the case for continuous exponential families, the slightest degree of mis-specification of the marginal distribution (see Section 4.2) or serial dependence structure (see Section 4.4) of arm rewards can cause optimized algorithms to lose the basic consistency property and suffer expected regret growing polynomially in the time horizon. Moreover, in such settings there is no control over any higher moment of the regret beyond the first moment (see Section 4.5). It is furthermore striking that every environment instance within such classes of bandit environments suffers from these fragility issues, not just some worst case instances within such classes.

Theorem 1 relies in part on the notion of *discrimination equivalence*, as stated in Definition 3 below. This property can be readily verified from (3). Following the statement of the theorem, we will provide an easier-to-verify equivalent characterization (Lemma 1) as well as simple sufficient conditions for this property (Propositions 1 and 2). We will then explain the choice of terminology, "discrimination equivalence", and provide examples for intuition.

Definition 3 (Discrimination Equivalence). A distribution P is *discrimination equivalent* if for any $z_1, z_2 \in \mathcal{I}_P$ with $z_1 > z_2$,

$$\inf_{z \in \mathcal{I}_P : z < z_2} \frac{d_P(z, z_1)}{d_P(z, z_2)} = 1. \tag{6}$$

For an algorithm π operating in an environment ν, we say that the resulting distribution of regret $R(T)$ has a *tail exponent* of $-c$ if $\mathbb{P}_{\nu\pi}(R(T) > x) \asymp x^{-c}$ as $T \to \infty$, uniformly in x with $T^a \leq x \leq a'T$, for any $0 < a < 1$ and suitable $a' > 0$. Intuitively, the regret tail exponent is determined by the tail exponent of the distribution of $N_{r(2)}(T)$, the number of plays of the second-best arm $r(2)$. (So it suffices to consider the tail exponent of $N_{r(2)}(T)$ when discussing the regret tail exponent.) In

Theorem 1, (7) and (8) are reflective of this intuition, since achieving logarithmic expected regret means the regret tail exponent cannot be greater than -1. (See Theorem 2 in Section 3.3, where we fully establish this intuition by specializing the analysis from general optimized algorithms to the KL-UCB algorithm.)

The full proof of Theorem 1 is given in Appendix A. In Section 3.2, we prove a simplified version of Theorem 1, along with a discussion to highlight the intuition behind this result. Through simulation studies (see Figures 1-3 in Section 7), we verify that the result provides accurate approximations over reasonably short time horizons.

THEOREM 1. *Let π be $\mathcal{M}_P$-optimized. Then for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ and the i-th-best arm $r(i)$,*

$$\liminf_{T\to\infty} \inf_{x\in B_\gamma(T)} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} \geq -\sum_{j=1}^{i-1} \inf_{z\in\mathcal{I}_P\,:\,z<\mu_{r(i)}} \frac{d_P(z,\mu_{r(j)})}{d_P(z,\mu_{r(i)})}, \tag{7}$$

with $B_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$ and any $\gamma \in (0,1)$.
If in addition, P is discrimination equivalent, then for the second-best arm $r(2)$,

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(2)}(T) > x)}{\log(x)} = -1 \tag{8}$$

uniformly for $x \in [T^\gamma, (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$. Moreover, for $i \geq 3$, (7) holds with the right side equal to $-(i-1)$.

In Lemma 1 below, we provide an equivalent characterization of discrimination equivalence. This characterization implies that each summand on the right side of (7) is equal to -1. (Note that for any $z, z_1, z_2 \in \mathcal{I}_P$ with $z < z_2 < z_1$, we always have $d_P(z,z_1)/d_P(z,z_2) \geq 1$.) In light of the exact -1 tail exponent for the second-best arm $r(2)$ in (8), we might conjecture that the lower bounds in (7) are tight in general without discrimination equivalence. We will rigorously establish this fact for a particular choice of algorithm (KL-UCB) in Section 3.3. The proof of Lemma 1 is given in Appendix EC.1.

LEMMA 1. *P is discrimination equivalent if and only if $\inf \Theta_P = -\infty$ and*

$$\lim_{\theta\to-\infty} \theta\Lambda_P'(\theta) - \Lambda_P(\theta) = \infty. \tag{9}$$

((9) can also be expressed as: $\lim_{\theta\to-\infty} \Lambda_P^*(\Lambda_P'(\theta)) = \infty$, where Λ_P^* is the convex conjugate of Λ_P.)

In Proposition 1, we give a simple sufficient condition for discrimination equivalence that applies to reward distributions with support that is unbounded to the left on the real line. The requirement is that the CGF of the distribution is finite on the negative half of the real line. In Proposition

2, we provide simple conditions to determine whether or not discrimination equivalence holds for distributions with support that is bounded to the left on the real line. When the support is bounded to the left, discrimination equivalence holds for continuous distributions, but generally not for discrete distributions. The proofs of Propositions 1 and 2 can be found in Appendix EC.1.

PROPOSITION 1. *If the support of P is unbounded to the left, and $\inf\Theta_P = -\infty$, then P is discrimination equivalent.*

PROPOSITION 2. *If the support of P is bounded to the left with no point mass at the infimum of the support, then P is discrimination equivalent. But if there is a positive point mass at the infimum of the support, then P is not discrimination equivalent.*

It can be verified that for fixed $z_1 > z_2$,

$$\inf_{z\in\mathcal{I}_P\,:\,z<z_2} \frac{d_P(z,z_1)}{d_P(z,z_2)} = \lim_{z\downarrow\inf\mathcal{I}_P} \frac{d_P(z,z_1)}{d_P(z,z_2)}. \tag{10}$$

The KL divergence $d_P(z,z')$ can be thought of as the mean information for discriminating between P^z and $P^{z'}$, given a sample from P^z. Since $d_P(z,z_1) = d_P(z,z_2)$ if and only if $z_1 = z_2$, the ratio $d_P(z,z_1)/d_P(z,z_2)$ can be thought of as a measure of the difficulty of discriminating between P^z and P^{z_1} relative to that between P^z and P^{z_2}, given a sample from P^z in both cases. The more similar the difficulty in the two cases, the closer the ratio is to one. In such cases, as suggested by Theorem 1, the regret tail will be heavier/closer to being truncated Cauchy. (In Theorem 2 in Section 3.3, we provide matching upper bounds for (7) for the KL-UCB algorithm, thereby providing validation for this way of thinking.) With this interpretation, we review in the following examples some of the settings covered by Propositions 1 and 2 above. We provide derivations for these examples at the end of Appendix EC.1.

EXAMPLE 1. Suppose in (1) that the base distribution P is the Gaussian distribution with mean 0 and variance σ^2. Then,

$$d_P(z,z') = \frac{(z-z')^2}{2\sigma^2}.$$

Hence, in this setting, (10) is always equal to 1, and P is discrimination equivalent.

EXAMPLE 2. Suppose in (1) that the base distribution P is the uniform distribution on $[0,1]$. Then, the CGF is:

$$\Lambda_P(\theta) = \log\left(\frac{e^\theta - 1}{\theta}\right).$$

It can be verified from the identity (3) that in this setting, (10) is always equal to 1, and so P is discrimination equivalent.

EXAMPLE 3. Suppose in (1) that the base distribution P is the Bernoulli distribution with mean 1/2. It can be verified from the identity (3) that in this setting,

$$\lim_{z\downarrow 0} \frac{d_P(z,z_1)}{d_P(z,z_2)} = \frac{\log(1-z_1)}{\log(1-z_2)}.$$

Hence, in this setting, (10) is always strictly greater than 1 for $0 < z_2 < z_1 < 1$, and so P is not discrimination equivalent.

A similar behavior arises whenever P puts positive mass at the left endpoint of its support, which we denote by L. From the perspective of the distribution P^z (which becomes a unit point mass at L as $z \downarrow L$), the different point masses at L associated with P^{z_1} and P^{z_2} can be discriminated at different rates. Hence, in such settings, P is not discrimination equivalent.

EXAMPLE 4. Suppose in (1) that the base distribution $P(dx) = e^x \cdot \mathbb{I}(x \le 0)\, dx$ for $x \in \mathbb{R}$, so P is a negatively supported exponential distribution, and $\Theta_P = (-1, \infty)$. It can be verified from the identity (3) that

$$\lim_{z\to-\infty} \frac{d_P(z,z_1)}{d_P(z,z_2)} = \frac{z_2}{z_1}.$$

Hence, in this setting, (10) is always strictly greater than 1 for $z_2 < z_1 < 0$, and so P is not discrimination equivalent. Intuitively, this behavior arises because P^{z_1} is a scale change of P^{z_2} (as opposed to a location change, as in the setting of Example 1). So the ability to discriminate from the perspective of P^z (as $z \to -\infty$), differs in the two cases, regardless of how negative z is.

As noted earlier, Theorem 1 establishes under P-discrimination equivalence that the regret tail of an $\mathcal{M}_P$-optimized algorithm is truncated Cauchy for every environment in $\mathcal{M}_P^K$. However, regardless of whether or not discrimination equivalence holds, there always exist some environments for which the regret tail of optimized algorithms is arbitrarily close to being truncated Cauchy (with a tail exponent arbitrarily close to -1). This is the content of Corollary 1 below, which follows immediately from (7) in Theorem 1 by taking the difference $\mu_{r(1)} - \mu_{r(2)}$ to be sufficiently small and using the relevant continuity property of the ratio of KL divergences on the right side of (7). This result highlights a universal fragility property of algorithms optimized for any exponential family class of environments. However, compared to the fragility implications from Theorem 1 which pertain to *all environment instances* within a class, Corollary 1 is weaker as it pertains only to *some environment instances* within a class.

COROLLARY 1. *Let π be $\mathcal{M}_P$-optimized. Then for any $\epsilon > 0$, there exists $\delta > 0$ such that for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ with $0 < \mu_{r(1)} - \mu_{r(2)} < \delta$,*

$$\liminf_{T\to\infty} \inf_{x\in B_\gamma(T)} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(2)}(T) > x)}{\log(x)} \ge -(1+\epsilon),$$

with $B_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$ and any $\gamma \in (0,1)$.

3.2. Key Ideas Behind Theorem 1 and Further Results

Below we provide a proof of a simplified version of Theorem 1, focusing on the two-armed bandit setting. As we will see, the key idea behind our proof is a change of measure argument in which the reward distribution of the optimal arm is tilted so that its mean becomes less than that of the sub-optimal arm. Then, within the new environment resulting from the change of measure, we require control over the number of plays of the new sub-optimal arm. Lemma 2 below provides such control through a weak law of large numbers (WLLN) for the number of sub-optimal arm plays of optimized algorithms. Lemma 2 follows immediately for optimized algorithms due to a "one-sided" WLLN in Theorem 4 in Section 5.2, which is developed for completely general (possibly non-exponential family) models. Moreover, Lemma 4 from Section 5.1 extends Lemma 2 to such general models. For further discussion, see Remark 3 in Section 5.2.

LEMMA 2. *Let π be $\mathcal{M}_P$-optimized. Then for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ and each sub-optimal arm i,*

$$\frac{N_i(T)}{\log(T)} \to \frac{1}{d_P(\mu_i, \mu_*(\nu))}$$

in $\mathbb{P}_{\nu\pi}$-probability as $T \to \infty$.

We will show the following simplified version of Theorem 1. Let $a \in (0,1)$, and $\nu = (P^{\mu_1}, P^{\mu_2}) \in \mathcal{M}_P^2$ such that (without loss of generality) $\mu_1 > \mu_2$, i.e., arm 1 is optimal in ν. For any $\mathcal{M}_P$-optimized algorithm π, we will first obtain:

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > aT)}{\log(T)} \geq - \inf_{z \in \mathcal{I}_P : z < \mu_2} \frac{d_P(z, \mu_1)}{d_P(z, \mu_2)}. \tag{11}$$

If additionally P is discrimination equivalent, then

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > aT)}{\log(T)} = -1. \tag{12}$$

Proof for (11) and (12). To obtain (11), consider a new environment $\widetilde{\nu} = (P^{\widetilde{\mu}_1}, P^{\mu_2}) \in \mathcal{M}_P^2$ with $\widetilde{\mu}_1 < \mu_2$, i.e., arm 1 is sub-optimal in $\widetilde{\nu}$. By a change of measure from ν to $\widetilde{\nu}$,

$$\mathbb{P}_{\nu\pi}\left(N_2(T) > aT\right) = \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(N_2(T) > aT\right) \underbrace{\prod_{t=1}^{N_1(T)} \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t))}_{:= L_T(\mu_1, \widetilde{\mu}_1)}\right]. \tag{13}$$

Note that

$$\log L_T(\mu_1, \widetilde{\mu}_1) = N_1(T) \cdot \frac{1}{N_1(T)} \sum_{t=1}^{N_1(T)} \log \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t)).$$

Under $\widetilde{\nu}$, by Lemma 2,

$$\frac{N_1(T)}{\log(T)} \to \frac{1}{d_P(\widetilde{\mu}_1, \mu_2)} \tag{14}$$

in $\mathbb{P}_{\widetilde{\nu}\pi}$-probability as $T \to \infty$. Under $\widetilde{\nu}$, by (14) and the WLLN,

$$\frac{1}{N_1(T)} \sum_{t=1}^{N_1(T)} \log \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t)) \to -d_P(\widetilde{\mu}_1, \mu_1) \tag{15}$$

in $\mathbb{P}_{\widetilde{\nu}\pi}$-probability as $T \to \infty$. The WLLN's (14) and (15) then imply that for $\epsilon > 0$,

$$\log L_T(\mu_1, \widetilde{\mu}_1) \geq -(1+\epsilon)\frac{d_P(\widetilde{\mu}_1, \mu_1)}{d_P(\widetilde{\mu}_1, \mu_2)} \log(T) \tag{16}$$

with $\mathbb{P}_{\widetilde{\nu}\pi}$-probability converging to 1 as $T \to \infty$. Since (under $\widetilde{\nu}$) $\mathbb{P}_{\widetilde{\nu}\pi}(N_2(T) > aT) \to 1$, using (13) and (16), we obtain:

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > aT)}{\log(T)} \geq -\frac{d_P(\widetilde{\mu}_1, \mu_1)}{d_P(\widetilde{\mu}_1, \mu_2)}. \tag{17}$$

Note that $\widetilde{\mu}_1$ is a free variable that we can optimize over, subject to the constraints: $\widetilde{\mu}_1 < \mu_2$ and $\widetilde{\mu}_1 \in \mathcal{I}_P$. Doing so yields (11). The right side of (11) equals -1 if P is discrimination equivalent. As noted in the Introduction, for an $\mathcal{M}_P$-optimized algorithm π,

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > aT)}{\log(T)} \leq -1.$$

So if π is $\mathcal{M}_P$-optimized and P is discrimination equivalent, we obtain (12). □

To obtain (11), the "optimal" change of measure from ν to $\widetilde{\nu}$ in (13) essentially involves sending $\widetilde{\mu}_1 \downarrow \inf \mathcal{I}_P$, which can be quite extreme. For example, under the conditions of Proposition 1, $\inf \mathcal{I}_P = -\infty$ and the optimal change of measure would involve sending the optimal arm 1 mean $\widetilde{\mu}_1 \to -\infty$. This suggests that the primary way that large regret arises is when the mean of the optimal arm 1 is under-estimated to be below that of the sub-optimal arm 2, likely due to receiving some unlucky rewards early on in the bandit experiment. Arm 1 is then mis-labeled as sub-optimal, and the mis-labeling is not corrected for a long time, resulting in large regret.

To obtain, for example, a regret of $O(T)$ when the optimal arm 1 is mis-labeled as sub-optimal, there effectively needs to be $O(\log(T))$ unusually low rewards from arm 1. The probability of such a scenario is exponential in the number of arm 1 plays. So the probability decays as an inverse power of T.

One might also consider a different change of measure, where the distribution of the sub-optimal arm 2 is tilted so that its mean is above that of the optimal arm 1. This corresponds to the scenario where the mean of arm 2 is over-estimated to be above that of arm 1, and so arm 2 is mis-labeled as optimal.

To obtain, for example, a regret of $O(T)$ when the sub-optimal arm 2 is mis-labeled as optimal, there effectively needs to be $O(T)$ unusually high rewards from arm 2. The probability of such a scenario is exponential in the number of arm 2 plays. So the probability decays exponentially with T.

To accompany Theorem 1, we show in Proposition 3 that large regret is not due to over-estimation of sub-optimal arm means, but must therefore be due to under-estimation of the optimal arm mean. The proof of Proposition 3 is given in Appendix EC.2. (Here, we use $\widehat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{s=1}^{N_i(t)} X_i(s)$ to denote the sample mean of arm i rewards up to time t.)

PROPOSITION 3. *Let π be $\mathcal{M}_P$-optimized. Then for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$, any sub-optimal arm i, and any $\epsilon > 0$,*

$$\lim_{T\to\infty} \mathbb{P}_{\nu\pi} \left(|\widehat{\mu}_i(T) - \mu_i| \le \epsilon \mid N_i(T) > x \right) = 1$$

uniformly for $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$.

It is straightforward to obtain results such as (11) and (12) in multi-armed settings. To obtain lower bounds on the distribution tail of the number of plays $N_{r(i)}(T)$ of arm $r(i)$ (the i-th-best arm, for $i \ge 2$), we tilt the reward distributions of arms $r(1), \ldots, r(i-1)$ so that their means become less than that of arm $r(i)$. We choose the new environment $\widetilde{\nu}$ with the new arm parameter values, so that arm $r(i)$ becomes the optimal arm. The change of measure from ν to $\widetilde{\nu}$ then results in the product of $i-1$ likelihood ratios corresponding to the arms $r(1), \ldots, r(i-1)$. Subsequently, each of the tilted parameter values for arms $r(1), \ldots, r(i-1)$ can be optimized separately to yield, for example, (7). We refer the reader to the full proof of Theorem 1 in Appendix A.

3.3. Tail Probability Upper Bounds

In Theorem 1 from Section 3.1, we developed a lower bound (7) for the distribution tail of the number of plays $N_{r(i)}(T)$ of the i-th-best arm $r(i)$ (for $i \ge 2$) by an optimized algorithm. In the presence of discrimination equivalence, we showed in (8) that the tail exponent for $R(T)$, as determined by $N_{r(2)}(T)$, is exactly equal to -1. The lower bound part of this result is obtained using (7) and discrimination equivalence. The upper bound part follows directly from Markov's inequality, as discussed in the Introduction.

However, when discrimination equivalence does not hold, the upper bound derived from Markov's inequality does not match the lower bounds. As part of Theorem 2, we develop refined upper bounds for the tail of $N_{r(i)}(T)$ for all $i \ge 2$, for the KL-UCB algorithm (Algorithm 2 and Theorem 1 of Cappé et al. (2013)). These refined upper bounds exactly match the lower bounds in (7), thereby providing strong evidence that the lower bounds in (7) are tight more generally, regardless of whether or not discrimination equivalence holds. The proof of Theorem 2 is given in Appendix B.

THEOREM 2. *Let π be $\mathcal{M}_P$-optimized KL-UCB. Then for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ and the i-th-best arm $r(i)$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{z\in\mathcal{I}_P : z<\mu_{r(i)}} \frac{d_P(z,\mu_{r(j)})}{d_P(z,\mu_{r(i)})} \tag{18}$$

uniformly for $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$.

From (18), we see that the tail exponents for the distributions of $N_{r(i)}(T)$, $i \geq 3$ are always strictly less than that of $N_{r(2)}(T)$. So $N_{r(2)}(T)$ determines the exponent of the distribution tail of the regret $R(T)$; see also Remark 1 below. Indeed, when P is discrimination equivalent, Lemma 1 implies that the right side of (18) is exactly $-(i-1)$ for $i \geq 3$, which can be compared to (8). Whenever P is not discrimination equivalent (for example, for all discrete distributions with support bounded to the left and strictly positive mass on the infimum of the support; see Proposition 2), the right side of (18) is always strictly less than -1 for the second-best arm $r(2)$. So the regret tail is always strictly lighter than truncated Cauchy in such settings. We confirm this fact for Bernoulli environments through numerical simulations in Figure 3 in Section 7.

This indicates that an algorithm optimized for (and operating within) an environment class $\mathcal{M}_P^K$, when P is a discrete distribution, is in general less fragile than when P is a continuous distribution. However, recall from Corollary 1 that regardless of whether the reward distributions are discrete or continuous, there always exist environments in $\mathcal{M}_P^K$ for which the regret tail is arbitrarily close to being truncated Cauchy. Optimized algorithms universally suffer from this weaker sense of fragility. In fact, as we will see in Section 5.2, this is a key characteristic of optimized algorithms that, together with our change of measure argument, leads to a new proof of a generalized version of the Lai-Robbins lower bound. (See Theorem 4 and Theorem 4.)

REMARK 1. In the setting of Theorem 2, the distribution tail of the regret $R(T)$, as determined by that of $N_{r(2)}(T)$, depends only on the top two arm reward distributions. (In this case, the KL divergences in (18) only involve $P^{\mu_{r(1)}}$ and $P^{\mu_{r(2)}}$.) In contrast, all sub-optimal arms contribute to the expected regret.

We also point out that (18) in Theorem 2 holds uniformly over a greater range $[\log^{1+\gamma}(T), (1-\gamma)T]$ than the range $[T^\gamma, (1-\gamma)T]$ of (8) in Theorem 1. As discussed in the Introduction, in reference to the CLT's for regret developed in Fan and Glynn (2022), the large deviations of regret correspond to deviations from the expected regret that are of order $\log(T)$. While we do not analyze deviations on such a scale in this paper, we do interpolate between the $\log(T)$ and poly-T regions by considering the poly-$\log(T)$ region of the regret tail. Since we simply relied on logarithmic expected regret and Markov's inequality in Theorem 1 to establish the upper bound part of (8),

there we could not make conclusions about the poly-log(T) regions. Here in Theorem 2, however, we perform careful analysis to establish a more informative upper bound, which gives us insight about the poly-log(T) regions.

In Sections 4 and 6, we will frequently use the KL-UCB algorithm and general UCB algorithms as examples to illustrate fragility issues and modifications to alleviate fragility issues. In Theorem 2 above, we characterized the regret tail of $\mathcal{M}_P$-optimized KL-UCB operating within environments from $\mathcal{M}_P^K$, i.e., the environment is well-specified. Later in Proposition 4, we develop a result for general UCB algorithms operating in essentially arbitrary environments, including mis-specified ones.

4. Illustrations of Fragility

4.1. Overview of Results

Throughout Section 4, we highlight several ways in which optimized algorithms are fragile. Whereas previously we developed regret tail characterizations for optimized algorithms in *well-specified* settings, we now consider *mis-specified* settings. By mis-specified, we mean that an algorithm π is designed (possibly optimized) for some class of environments $\mathcal{M}^K$, but π operates in an environment $\nu \notin \mathcal{M}^K$. In real world settings, there is generally some degree of model mis-specification. So it is important to understand the sensitivity of algorithmic performance to mis-specification, of which there are many different forms.

Recall that in the formulation of instance-based asymptotic optimality of bandit algorithms, we first restrict consideration to the class of consistent algorithms, as in Definition 1. Then, within the class of consistent algorithms, we seek optimized algorithms that achieve the Lai-Robbins lower bound, as in Definition 2. From (8) in Theorem 1, in every (well-specified) environment instance, optimized algorithms have extremely heavy regret tails, essentially truncated Cauchy with tail exponent exactly equal to -1 (under the discrimination equivalence condition). If the tail exponent is at all > -1, then expected regret grows polynomially in the time horizon, i.e., consistency is lost. Thus, in every single (well-specified) environment instance, optimized algorithms just barely maintain consistency. Moreover, this suggests that the slightest degree of model mis-specification can cause optimized algorithms to not only suffer greater expected regret, but altogether lose the basic consistency property by suffering expected regret growing polynomially in the time horizon.

Our results in Sections 4.2-4.4 indicate that the hypothesis of the previous sentence is true for optimized algorithms in great generality. It is useful to highlight two particular examples involving KL-UCB (an optimized algorithm in well-specified settings), which were touched upon in the Introduction and will be discussed in more detail in the sections to follow. First, consider KL-UCB designed for iid Gaussian rewards with variance σ^2, and operating in an environment with iid

Gaussian rewards with variance $\sigma_0^2 > \sigma^2$. Second, consider the same KL-UCB algorithm designed for iid Gaussian rewards, but operating in an environment with rewards evolving according to a stationary AR(1) process with positive AR coefficient and Gaussian marginal distributions matching the algorithm's Gaussian design (with the same variance σ^2). So the first case corresponds to mis-specified marginal distributions, whereas the second case corresponds to mis-specified correlation structure. In both cases, no matter how close σ_0^2 is to σ^2, or how close the AR(1) coefficient is to zero, the KL-UCB algorithm will have a regret tail that is heavier than truncated Cauchy, with tail exponent > -1. Thus, KL-UCB is inconsistent in these settings, with expected regret growing polynomially in the time horizon. Strikingly, these results hold no matter how big the separation is between the arm means.

In Section 4.2, we study mis-specification of the marginal distributions of rewards in iid settings. Then, in Section 4.3, we develop lower bounds on the regret tail for general reward processes, which are subsequently applied to study mis-specification of the serial dependence structure of rewards in Section 4.4. Our theory in these sections will primarily be developed for $\mathcal{M}_P$-optimized KL-UCB for any chosen base distribution P, and operating in an environment $\nu \notin \mathcal{M}_P^K$. To avoid pathological/trivial situations and ensure that the KL divergence function d_P remains well-defined in mis-specified settings, we will assume that $\mathcal{I}_P$ is an interval (possibly infinite) that contains the range of all possible values of rewards for each arm of the true environment ν.

To conclude our illustrations of fragility, we examine in Section 4.5 the higher moments (beyond the first moment) of regret for optimized algorithms operating in well-specified environments. Under the assumptions of Theorem 1, we will see that optimizing for expected regret provides no control (uniform integrability) over any higher power of regret. Higher moments grow as powers of the time horizon T instead of as poly-log(T).

4.2. Mis-specified Reward Distribution

In this section, we examine the regret tail behavior of optimized algorithms under mis-specification of marginal reward distributions. We begin with Proposition 4, which is a characterization of the regret tail of (possibly) mis-specified KL-UCB operating in an environment $\nu = (Q_1, \ldots, Q_K)$, where arm i yields independent rewards from some distribution Q_i. We can compare the right side of (19) in Proposition 4 to the right side of (18) in Theorem 2. In well-specified settings, Theorem 2 and Proposition 4 are the same result. In mis-specified settings, which is covered by Proposition 4, the KL divergences $d_{Q_{r(j)}}$ in the numerator do not match the KL divergence d_P in the denominator.

The proof of Proposition 4 is given in Appendix EC.3. The proof uses a LLN (Lemma 3) for the regret of (possibly) mis-specified KL-UCB, and a general tail probability lower bound (Theorem 3), which are deferred to Section 4.3. These supporting results are developed for more general

(possibly non-iid) reward processes. They are useful for establishing the results in Section 4.4, but they are stronger than needed in the current section.

PROPOSITION 4. *Let π be $\mathcal{M}_P$-optimized KL-UCB. Let the environment $\nu = (Q_1, \ldots, Q_K)$, where arm i yields independent rewards from some distribution Q_i such that its CGF $\Lambda_{Q_i}(\theta) < \infty$ for θ in a neighborhood of zero. Then for the i-th-best arm $r(i)$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{z\in\mathcal{I}_{Q_{r(j)}} : z<\mu(Q_{r(i)})} \frac{d_{Q_{r(j)}}(z, \mu(Q_{r(j)}))}{d_P(z, \mu(Q_{r(i)}))} \tag{19}$$

uniformly for $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$.

In Corollary 2 below, we show that for Gaussian KL-UCB operating in environments with iid Gaussian rewards, if the actual variance is just slightly greater than the variance specified in the algorithm design, then the expected regret will grow at a rate that is a power of T. The proof details simplify significantly in this Gaussian setting, and for future reference, we provide a stand-alone proof of Corollary 2 in Appendix EC.3. See Figure 1 in Section 7 for numerical simulations illustrating (20).

COROLLARY 2. *Let π be KL-UCB optimized for iid Gaussian rewards with variance $\sigma^2 > 0$. Then for any two-armed environment ν yielding iid Gaussian rewards with actual variance $\sigma_0^2 > 0$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(2)}(T) > x)}{\log(x)} = -\frac{\sigma^2}{\sigma_0^2} \tag{20}$$

uniformly for $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$. So if $\sigma_0^2 > \sigma^2$, then for any $a \in (\sigma^2/\sigma_0^2, 1]$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_{r(2)}(T)]}{T^{1-a}} \geq 1.$$

Corollary 2 also holds with π as TS designed for iid Gaussian rewards with variance $\sigma^2 > 0$ (and with Gaussian priors on the arm means). In particular, we can obtain this result by using the SLLN's developed in Fan and Glynn (2022).

4.3. Tail Probability Lower Bounds for General Reward Processes

In this section, we develop supporting results, which are needed in Section 4.4 to establish regret tail characterizations in settings where the dependence structures of rewards are mis-specified. These supporting results can also be used to derive the results in Section 4.2 in settings where the marginal reward distributions are mis-specified. Lemma 3 is a SLLN for the regret of KL-UCB operating in an environment with general (possibly non-iid) reward processes that satisfy Assumptions 1-2 below. KL-UCB is an example of an algorithm that is so-called $\mathcal{M}_P$-*pathwise convergent*, a notion

that we introduce in Definition 4 below. In Theorem 3, we apply our change-of-measure argument to establish lower bounds for the regret tail of such algorithms when operating in an environment with reward processes satisfying Assumptions 1-2.

We first state a few definitions and assumptions for the reward processes $X_i(t)$, $i \in [K]$, $t \geq 1$ that we will work with. For each arm i and sample size $n \geq 1$, define the re-scaled CGF of the sample mean of arm rewards:

$$\overline{\Lambda}_i^n(\theta) = \frac{1}{n} \log \mathbb{E}\left[\exp\left(\theta \cdot \sum_{t=1}^{n} X_i(t)\right)\right].$$

We will assume the following for each arm i.

ASSUMPTION 1. *The limit* $\overline{\Lambda}_i(\theta) = \lim_{n\to\infty} \overline{\Lambda}_i^n(\theta)$ *exists (possibly infinite) for each* $\theta \in \mathbb{R}$, *and* $0 \in \overline{\Theta}_i := \text{interior}\{\theta \in \mathbb{R} : \overline{\Lambda}_i(\theta) < \infty\}$.

ASSUMPTION 2. $\overline{\Lambda}_i(\cdot)$ *is differentiable throughout* $\overline{\Theta}_i$, *and either* $\overline{\Theta}_i = \mathbb{R}$ *or* $\lim_{m\to\infty} \left|\overline{\Lambda}_i'(\theta^m)\right| = \infty$ *for any sequence* $\theta^m \in \overline{\Theta}_i$ *converging to a boundary point of* $\overline{\Theta}_i$.

These are the conditions ensuring that the Gärtner-Ellis Theorem holds for the sample means of arm rewards (see, for example, Theorem 2.3.6 of Dembo and Zeitouni (1998)). In the context of Assumption 1, we refer to the limit $\overline{\Lambda}_i$ as the *limiting CGF* for arm i. In the context of Assumption 2, $\overline{\Lambda}_i'(0)$, the derivative of limiting CGF evaluated at zero, is the long-run mean reward for arm i. Indeed, by the Gärtner-Ellis Theorem and the Borel-Cantelli Lemma,

$$\frac{1}{n}\sum_{t=1}^{n} X_i(t) \to \overline{\Lambda}_i'(0) \tag{21}$$

almost surely as $n \to \infty$ for each arm i. The optimal arm $r(1)$ is such that $\overline{\Lambda}_{r(1)}'(0) = \max_{i\in[K]} \overline{\Lambda}_i'(0)$.

In the current section and in Section 4.4, we also assume for simplicity that the reward process for each arm only evolves forward in time when the arm is played. This ensures that the serial dependence structures of the reward processes are not interrupted in a complicated way by an algorithm's adaptive sampling schedule, and allows us to determine the limit in Assumption 1 for various processes of interest such as Markov chains. Regardless of the specifics of the serial dependence structure of rewards for each arm, we will always assume that there is no dependence between rewards of different arms.

Before stating Lemma 3 and Theorem 3, we introduce the following notion, which can be compared to the notion of an $\mathcal{M}_P$-optimized algorithm in Definition 2.

DEFINITION 4 (PATHWISE CONVERGENT ALGORITHM). An algorithm π is $\mathcal{M}_P$*-pathwise convergent* if for any environment ν yielding arm reward sequences $X_i(t)$, $i \in [K]$, $t \geq 1$,

$$\left\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} X_i(t) = c_i,\ i \in [K]\right\} \subset \left\{\omega : \lim_{T\to\infty} \frac{N_i(T)}{\log(T)} = \frac{1}{d_P(c_i, \max_j c_j)},\ \forall\, i \neq \arg\max_j c_j\right\}. \tag{22}$$

We conjecture that, in general, $\mathcal{M}_P$-optimized algorithms are also $\mathcal{M}_P$-pathwise convergent. This is directly supported by Lemma 3 below, as well as by the SLLN developed for Gaussian TS in Fan and Glynn (2022). It is also suggested by the analysis for developing SLLN's for non-optimized forced sampling-based algorithms and other UCB algorithms in Cowan and Katehakis (2019). The proof of Lemma 3 is based on the arguments in Cowan and Katehakis (2019), and can be found in Appendix EC.4.

LEMMA 3. *$\mathcal{M}_P$-optimized KL-UCB is $\mathcal{M}_P$-pathwise convergent.*

We now introduce Theorem 3, whose proof can be found in Appendix C. For arm i, we use $\overline{\Lambda}_i^*$ to denote the convex conjugate of the limiting CGF $\overline{\Lambda}_i$, and we define $\overline{\mathcal{I}}_i = \text{interior}\{z \in \mathbb{R} : \overline{\Lambda}_i^*(z) < \infty\}$. As mentioned previously, to avoid pathological/trivial situations, we will always assume for each arm i that $\overline{\mathcal{I}}_i \subset \mathcal{I}_P$ for the chosen base distribution P. (We also recall that the convex conjugate of the limiting CGF is the *rate function* in the Gärtner-Ellis Theorem.)

THEOREM 3. *Let π be $\mathcal{M}_P$-pathwise convergent. Let the K-armed environment ν yield rewards for each arm that evolve according to any process satisfying Assumptions 1-2. Then for the i-th-best arm $r(i)$,*

$$\liminf_{T\to\infty} \inf_{x\in B_\gamma(T)} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} \geq -\sum_{j=1}^{i-1} \inf_{z\in\overline{\mathcal{I}}_{r(j)} : z<\overline{\Lambda}'_{r(i)}(0)} \frac{\overline{\Lambda}^*_{r(j)}(z)}{d_P(z, \overline{\Lambda}'_{r(i)}(0))}, \tag{23}$$

with $B_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$ and any $\gamma \in (0,1)$.

REMARK 2. Whenever we can establish a WLLN for the $N_i(T)$ (e.g., as in Lemma 2), then our change-of-measure approach can be used to obtain lower bounds on the tail probabilities of the $N_i(T)$ (as in Theorems 1 and 3). The almost sure convergence of the $N_i(T)$, as provided by Assumptions 1-2 (leading to (21)) together with pathwise convergence (in Definition 4), is sufficient but not necessary.

4.4. Mis-specified Reward Dependence Structure

Even if the marginal distributions of the arm rewards are correctly specified, optimized algorithms such as KL-UCB (designed for iid rewards) can still be susceptible to mis-specification of the serial dependence structure. In Corollary 3, we provide a lower bound characterization of the regret tail for Gaussian KL-UCB applied to bandits with rewards evolving as Gaussian AR(1) processes. Specifically, for each arm i, we assume the rewards evolve as an AR(1) process:

$$X_i(t) = \alpha_i + \beta_i X_i(t-1) + W_i(t), \tag{24}$$

where the $\beta_i \in (0,1)$ and the $W_i(t)$ are iid $N(0,\sigma_i^2)$. The equilibrium distribution for arm i is then $N(\alpha_i/(1-\beta_i), \sigma_i^2/(1-\beta_i^2))$. For simplicity, we assume that the AR(1) reward process for each arm is initialized in equilibrium. So the marginal mean (also the long-run mean as in (21)) for arm i is $\overline{\Lambda}_i'(0) = \alpha_i/(1-\beta_i)$. The proof of Corollary 3 follows from a straightforward verification of Assumptions 1-2, which is omitted, and then a direct application of Theorem 3.

COROLLARY 3. *Let π be KL-UCB optimized for iid Gaussian rewards with variance $\sigma^2 > 0$. Then for any two-armed environment ν yielding rewards that evolve as AR(1) processes (as in (24)),*

$$\liminf_{T\to\infty} \inf_{x\in B_\gamma(T)} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(2)}(T) > x)}{\log(x)} \geq -\frac{\sigma^2}{\sigma_{r(1)}^2}(1-\beta_{r(1)})^2,$$

with $B_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$ and any $\gamma \in (0,1)$.

To see the effect of mis-specifying the dependence structure, suppose $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$ and $\beta_1 = \beta_2 = \beta_0$, for some $\sigma_0^2 > 0$ and $\beta_0 \in (0,1)$, so that the equilibrium distributions for the rewards of both arms are Gaussian with variance $\sigma_0^2/(1-\beta_0^2)$. Then, even if we specify the same variance $\sigma^2 = \sigma_0^2/(1-\beta_0^2)$ in Gaussian KL-UCB, so that the marginal distribution of rewards is correctly specified, we still end up with a tail exponent that is strictly greater than -1. This is due to the mis-specification of the serial dependence structure. Specifically, using Corollary 3,

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(2)}(T) > T/2)}{\log(T)} \geq -\frac{1-\beta_0}{1+\beta_0}, \tag{25}$$

and so for any $a \in ((1-\beta_0)/(1+\beta_0), 1]$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_{r(2)}(T)]}{T^{1-a}} \geq 1.$$

We verify (25) through numerical simulations in Figure 2 in Section 7. The simulations suggest that the lower bound in (25) is tight.

In Proposition 5 below, we develop a characterization of the regret tail of KL-UCB operating in an environment ν with rewards evolving as finite state Markov chains. For each arm i, we assume that the rewards evolve as an irreducible Markov chain on a common, finite state space $S \subset \mathbb{R}$, with transition matrix H_i. For any $\theta \in \mathbb{R}$ and transition matrix H, we use $\phi_H(\theta)$ to denote the logarithm of the Perron-Frobenius eigenvalue of the corresponding tilted transition matrix:

$$\big(\exp(\theta \cdot y) H(x,y),\ x,y \in S\big). \tag{26}$$

So in the context of Assumptions 1-2, $\overline{\Lambda}_i(\theta) = \phi_{H_i}(\theta)$ for each arm i. (Note that the convex conjugate $\overline{\Lambda}_i^*$ of $\overline{\Lambda}_i$ plays the same role in Proposition 5 as it does in Theorem 3.) For simplicity, we assume that the Markov chain reward process for each arm is initialized in equilibrium. So the marginal

mean (also the long-run mean as in (21)) for arm i is $\overline{\Lambda}_i'(0) = \phi'_{H_i}(0)$. Lastly, we wish to ensure that any equilibrium mean between $s_{\min} := \min S$ and $s_{\max} := \max S$ can be realized through tilting the transition matrices as in (26). This provides technical convenience, and allows us to use Chernoff-type bounds for Markov chains from the existing literature to derive upper bounds on the regret tail. So we introduce the following notion. We say that a transition matrix H on S satisfies the *Doeblin Condition* if we have $H(x, s_{\min}) > 0$ for each $x \neq s_{\min}$, and $H(x, s_{\max}) > 0$ for each $x \neq s_{\max}$.

The lower bound part of Proposition 5 follows from a straightforward verification of Assumptions 1-2, which is omitted, and then a direct application of Theorem 3. To establish the upper bound part, we can again use the proof of Theorem 2 (in Appendix B) and substitute in, where appropriate (in (60) and (65)), a Chernoff-type bound for additive functionals of finite-state Markov chains. One version of such a result that is convenient for our purposes is established in Theorem 1 of Moulos and Anantharam (2019). (Earlier and more general results can be found in Miller (1961) and Kontoyiannis and Meyn (2003), respectively.)

PROPOSITION 5. *Let π be $\mathcal{M}_P$-optimized KL-UCB. Let the K-armed environment ν yield rewards for each arm that evolve according to an irreducible Markov chain with a finite state space (with $\overline{\Lambda}_i$ as defined above for each arm i), and suppose that the transition matrix for each arm satisfies the Doeblin Condition. Then for the i-th-best arm $r(i)$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{z\in\overline{\mathcal{I}}_{r(j)} : z<\overline{\Lambda}'_{r(i)}(0)} \frac{\overline{\Lambda}^*_{r(j)}(z)}{d_P(z, \overline{\Lambda}'_{r(i)}(0))} \tag{27}$$

uniformly for $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$.

EXAMPLE 5. For the state space $S = \{0,1\}$ (binary rewards), we can examine some numerical values for the right side of (27). Here, we take $d_P(z,z')$ to be the KL divergence between Bernoulli distributions with means z and z'. We assume the arm rewards evolve as Markov chains on S. So the marginal distributions of the arm rewards are well-specified. Suppose the best arm $r(1)$ evolves according to a transition matrix of the form:

$$H_{r(1)} = \begin{bmatrix} 1-q & q \\ 1-w(q) & w(q) \end{bmatrix}. \tag{28}$$

For any $q \leq 0.8$, we set $w(q) \geq 0.8$ such that the chain evolving on S according to $H_{r(1)}$ has equilibrium mean equal to 0.8. Suppose also that the gap between the equilibrium means of the top two arms, $r(1)$ and $r(2)$, is $\Delta > 0$. In Table 1 below, we provide numerical values for the right side of (27) for the case $i = 2$ and for different values of q and Δ. As q becomes smaller relative to 0.8, the autocorrelation in the rewards for arm $r(1)$ becomes more positive, and the resulting regret distribution tail becomes heavier. As the gap Δ shrinks, the resulting regret tail also becomes

heavier. We can see from Table 1 that it is fairly easy (for reasonable values of q and Δ) to obtain regret tails that are heavier than truncated Cauchy (the right side of (27) is greater than -1).

		Δ								
q	$w(q)$	0.12	0.11	0.10	0.09	0.08	0.07	0.06	0.05	0.04
0.8	0.8	-1.41	-1.37	-1.34	-1.30	-1.26	-1.23	-1.19	-1.16	-1.13
0.7	0.825	-1.06	-1.03	-1.00	-0.97	-0.95	-0.92	-0.89	-0.87	-0.84
0.6	0.85	-0.80	-0.78	-0.76	-0.74	-0.72	-0.70	-0.68	-0.66	-0.64
0.5	0.875	-0.61	-0.59	-0.58	-0.56	-0.54	-0.53	-0.51	-0.50	-0.49

Table 1 For arm rewards evolving as Markov chains on the state space $S=\{0,1\}$, and with d_P being the Bernoulli KL divergence, we provide numerical values for the right side of (27). Here, $i=2$, and we consider different values of q (as used in the best arm's transition matrix $H_{r(1)}$ in (28)) and Δ (the gap between the equilibrium means of the two best arms, $r(1)$ and $r(2)$).

4.5. Higher Moments

In this section, we point out that the $1+\delta$ moment of regret for any $\delta>0$ must grow roughly as T^{δ}. Contrary to what one might conjecture in light of the WLLN that we saw in Lemma 2, the $1+\delta$ moment of regret is not poly-logarithmic. In Corollary 4 below, which is a direct consequence of Theorem 1, we show that expected regret minimization does not provide any help in controlling higher moments of regret. It forces the tail of the regret distribution to be as heavy as possible while ensuring the expected regret scales as $\log(T)$ (as we saw in Theorem 1 and Corollary 1). Consequently, there is no control over the distribution tails of $1+\delta$ powers of regret, and thus no uniform integrability of $1+\delta$ powers of regret (normalized by $\log^{1+\delta}(T)$).

COROLLARY 4. *Let π be $\mathcal{M}_P$-optimized. Suppose also that P is discrimination equivalent. Then for any environment $\nu\in\mathcal{M}_P^K$, and any $\delta>0$ and $\delta'\in(0,\delta)$,*

$$\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu\pi}[N_{r(2)}(T)^{1+\delta}]}{T^{\delta'}}\geq 1.$$

5. A Generalization and a Trade-off

In this section, we consider well-specified settings and develop further results regarding the regret tail. We consider a general model denoted by $\mathcal{M}$, which is allowed to be an arbitrary collection of distributions with finite means. In Section 5.1, Proposition 6 is a lower bound for the regret tail for such general models $\mathcal{M}$, which generalizes the lower bound in (7) from Theorem 1 (developed for exponential family models). In Section 5.2, in the context of a general model $\mathcal{M}$, Theorem 4 exposes a trade-off between the heaviness of the regret tail and the growth rate of expected regret, with lighter regret tails implying larger growth rates of expected regret.

5.1. Tail Probability Lower Bounds for General Models

Before establishing Proposition 6, which is a lower bound for the regret tail of optimized algorithms in a general model $\mathcal{M}$ (to be defined below), we first discuss some concepts. We will use the following quantity frequently:

$$D_{\inf}(Q, y, \mathcal{M}) := \inf\{D(Q \,\|\, Q') : Q' \in \mathcal{M},\ \mu(Q') > y\}, \tag{29}$$

which is defined for a general model $\mathcal{M}$, an arbitrary distribution Q, and $y \in \mathbb{R}$. Here, $D(Q \,\|\, Q')$ denotes the KL divergence between the distributions Q and Q', and we take the infimum of the empty set to be $+\infty$. When $\mathcal{M} = \mathcal{M}_P$, an exponential family with base distribution P, then $D_{\inf}(P^\mu, \mu', \mathcal{M}_P) = d_P(\mu, \mu')$ for $\mu < \mu'$.

Next, we have the definition of optimized algorithm for a general model $\mathcal{M}$ (similar to Defintion 2 for exponential family models). This notion of optimized algorithm comes from the generalized version (due to Burnetas and Katehakis (1996)) of the Lai-Robbins lower bound for expected regret (from (5)).

Definition 5 ($\mathcal{M}$-Optimized Algorithm). For a general model $\mathcal{M}$, an algorithm π is $\mathcal{M}$-optimized if for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and each sub-optimal arm i,

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{\log(T)} = \frac{1}{D_{\inf}(P_i, \mu_*(\nu), \mathcal{M})}.$$

The above definition of optimized algorithm leads to the regret tail lower bound in Proposition 6 below. This result generalizes the regret tail lower bound in (7) from Theorem 1. Its proof, which we omit, follows from a direct adaptation of the proof of (7) from Theorem 1 (see Appendix A). As in the proof of (7) from Theorem 1, we require a WLLN for the number of sub-optimal arm plays of $\mathcal{M}$-optimized algorithms. This is provided in Lemma 4 below, which is an extension of the earlier Lemma 2 to general models $\mathcal{M}$. The WLLN in Lemma 4 follows immediately from (32) in Theorem 4 with the choice $f(T) = \log(T)$ (as discussed in Remark 3 in the next section), together with the definition of an $\mathcal{M}$-optimized algorithm.

Lemma 4. *Let π be $\mathcal{M}$-optimized. Then for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and each sub-optimal arm i,*

$$\frac{N_i(T)}{\log(T)} \to \frac{1}{D_{\inf}(P_i, \mu_*(\nu), \mathcal{M})}$$

in $\mathbb{P}_{\nu\pi}$-probability as $T \to \infty$.

PROPOSITION 6. *Let π be $\mathcal{M}$-optimized. Then for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and the i-th best arm $r(i)$,*

$$\liminf_{T\to\infty} \inf_{x\in B_\gamma(T)} \frac{\mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} \geq -\sum_{j=1}^{i-1} \inf_{Q\in\mathcal{M}\,:\,\mu(Q)<\mu(P_{r(i)})} \frac{D(Q \,\|\, P_{r(j)})}{D_{\inf}(Q, \mu(P_{r(i)}), \mathcal{M})}, \tag{30}$$

with $B_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$ and any $\gamma \in (0,1)$.

When Proposition 6 is specialized to a model $\mathcal{M}_P$ as in (1)-(2), we recover (7) in Theorem 1. In this case, the infima of the individual optimization problems on the right side of (30) are attained or approached by taking $Q = P^z$ with $z \downarrow \inf \mathcal{I}_P$. In other words, the optimal change of measure is to exponentially tilt the mean of the optimal arm to be as small as possible; recall the discussion following (10) in Section 3.1. In practical terms, as discussed in Section 3.2, this suggests that large regret arises due to the mean of the optimal arm being severely under-estimated to be below the means of the sub-optimal arm(s). However, for general models $\mathcal{M}$, the distribution(s) $Q \in \mathcal{M}$ that attain or approach the infima for each of the individual optimization problems on the right side of (30) may not be straightforward to determine. In general, the optimal change of measure will not be a simple exponential tilt to change the mean.

We can already see interesting distinctions in the following example, where $\mathcal{M}$ is the set of all Gaussian distributions with unknown means and also unknown variances. Note that this is a strictly larger model than any model $\mathcal{M}_P$ with Gaussian base distribution P having a particular variance that is known (so that $\mathcal{M}_P$ is parameterized only by the unknown means).

EXAMPLE 6. Let $\mathcal{M}$ be the collection of all Gaussian distributions with all possible means and variances. This model corresponds to the setting with Gaussian rewards, where both the mean and variance are unknown. For any distribution $Q = N(z, \sigma^2)$ and $z < y$,

$$D_{\inf}(Q, y, \mathcal{M}) = \frac{1}{2} \log\left(1 + \frac{(y-z)^2}{\sigma^2}\right).$$

Let $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$, with $P_k = N(\mu_k, \sigma_k^2)$ and $\sigma_k^2 > 0$. Then, for $i = 2$ and $j = 1$, the optimization problem on the right side of (30) can be expressed as:

$$-\inf_{Q\in\mathcal{M}\,:\,\mu(Q)<\mu(P_{r(2)})} \frac{D(Q \,\|\, P_{r(1)})}{D_{\inf}(Q, \mu(P_{r(2)}), \mathcal{M})} = -\inf_{(z,\sigma^2)\,:\,z<\mu_{r(2)},\,\sigma^2>0} \frac{\log\left(\frac{\sigma_{r(1)}^2}{\sigma^2}\right) + \frac{\sigma^2 + (\mu_{r(1)} - z)^2}{\sigma_{r(1)}^2} - 1}{\log\left(1 + \frac{(\mu_{r(2)} - z)^2}{\sigma^2}\right)} = -1.$$

So, the regret tail in this setting is also truncated Cauchy. Here, the optimal value (tail exponent) of -1 is achieved by fixing any $z < \mu_{r(2)}$ and then sending $\sigma^2 \downarrow 0$. In other words, an optimal change of measure is to tilt the mean of the best arm to be any finite amount lower than the mean of the second-best arm, and then send the variance of the best arm to zero. By contrast, for a Gaussian

model $\mathcal{M}_P$ parameterized by mean with known variance (which is strictly smaller than $\mathcal{M}$), recall that the optimal change of measure is to tilt the mean of the best arm all the way to $-\infty$. This type of change of measure (a simple exponential tilting) is inadequate for achieving the -1 tail exponent when the model is $\mathcal{M}$.

5.2. Trade-off: Regret Tail vs. Regret Expectation

In this section, we establish a trade-off between the heaviness (decay rate) of the regret tail and the growth rate of expected regret. In particular, an upper bound on the decay rate of the regret tail implies a lower bound on the growth rate of expected regret. And a greater decay rate of the regret tail in (31) must be accompanied by a greater growth rate of expected regret. This holds for a range of different rates as captured through the choice of function f, from logarithmic to polynomial (in T), with the latter corresponding to exponential regret tails.

The trade-off we establish generalizes the asymptotic lower bounds for expected regret, first developed by Lai and Robbins (1985) for exponential families $\mathcal{M}_P$ and then extended by Burnetas and Katehakis (1996) to general models $\mathcal{M}$. To see this, take $f(T) = \log(T)$ in Theorem 4. Then, the assumption of consistency of algorithms used in these papers implies the regret tail upper bound in (31), which is the starting assumption of Theorem 4. More specifically, Lai and Robbins (1985) assumes $\mathcal{M}_P$-consistency, as in Definition 1 for exponential family models, and Burnetas and Katehakis (1996) assumes $\mathcal{M}$-consistency, which is defined in the same way except for general models $\mathcal{M}$. These consistency assumptions ensure, by Markov's inequality, that the regret tail cannot be heavier than truncated Cauchy in all such environments, i.e., the tail exponent must be ≤ -1, which is precisely the content of (31) for the choice $f(T) = \log(T)$.

We will study the tightness of the trade-off in Theorem 4 by analyzing an extension of the KL-UCB algorithm (see Algorithm 1) in Section 6.1 below. Specifically, see Proposition 7 and Remark 4. As will be discussed, a sufficient condition for the trade-off to be tight is when f is a so-called slowly varying function (intuitively, with growth slower than any polynomial). When f exhibits faster growth (for example, polynomial), then the lower bound for expected regret in (33) still has the correct asymptotic dependence on T, but the resulting constant on the right side may not be tight.

The proof of Theorem 4 can be found in Appendix D. The environments in the general environment class $\mathcal{M}^K$ are allowed to be arbitrary (provided the distributions have finite mean). The best arm(s), second-best arm(s), etc., do not need to be unique.

THEOREM 4. *Let $f:(1,\infty)\to(0,\infty)$ be an increasing function satisfying $\liminf_{t\to\infty} f(t)/\log(t)\geq 1$ and $f(t)=o(t)$. For every environment $\nu=(P_1,\ldots,P_K)\in\mathcal{M}^K$, suppose π satisfies for each sub-optimal arm i:*

$$\limsup_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_i(T)>T/K)}{f(T)}\leq -1. \tag{31}$$

Then for any such environment $\nu=(P_1,\ldots,P_K)\in\mathcal{M}^K$, each sub-optimal arm i, and any $\epsilon\in(0,1)$,

$$\lim_{T\to\infty}\mathbb{P}_{\nu\pi}\left(\frac{N_i(T)}{f(T)}\geq\frac{1-\epsilon}{D_{\inf}(P_i,\mu_*(\nu),\mathcal{M})}\right)=1, \tag{32}$$

and thus,

$$\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{f(T)}\geq\frac{1}{D_{\inf}(P_i,\mu_*(\nu),\mathcal{M})}. \tag{33}$$

REMARK 3. To see that Theorem 4, with the choice $f(T)=\log(T)$, implies Lemma 4, note the following. For $\mathcal{M}$-optimized algorithms satisfying Definition 5, (31) is automatically satisfied, since the regret tail of such algorithms (which are, of course, $\mathcal{M}$-consistent) cannot be heavier than truncated Cauchy, as discussed above. Thus, the "one-sided" WLLN in (32) holds for $\mathcal{M}$-optimized algorithms, which implies the full "two-sided" WLLN in Lemma 4. In the special case that $\mathcal{M}=\mathcal{M}_P$ is an exponential family, these arguments apply verbatim to justify Lemma 2 for $\mathcal{M}_P$-optimized algorithms satisfying Definition 2.

6. Improvement of the Regret Tail

In Sections 3 and 4, we have seen that the regret tails of algorithms optimized for restrictive model classes like exponential families are extremely heavy, to the point that such optimized algorithms are barely able to maintain consistency in well-specified settings. In particular, the slightest degree of model mis-specification can result in the loss of consistency, with expected regret growing polynomially in the time horizon. In Section 6.1, we discuss a simple approach to make the regret tail lighter for UCB-type algorithms. Our analysis establishes an explicit trade-off between the amount of UCB exploration (and corresponding expected regret) and the resulting heaviness of the regret tail (see Proposition 7, which also validates and strengthens the trade-off in Theorem 4). Moreover, we show in Section 6.2 that the modification to obtain lighter regret tails provides protection against mis-specification of the arm reward distributions in iid settings. We show in Section 6.3 that the modification also provides protection against Markovian departures from independence of the arm rewards.

6.1. A Simple Approach to Obtain Lighter Regret Tails

In this section, and in Sections 6.2 and 6.3, we focus on extensions of the KL-UCB algorithm of Cappé et al. (2013); see Algorithm 1 below. Like KL-UCB, Algorithm 1 is defined for any exponential family $\mathcal{M}_P$, as in (1)-(2). However, in contrast to KL-UCB, we will allow for more flexibility in the rate of exploration, as specified by the "exploration" function f. Whereas KL-UCB is developed for certain f satisfying $\lim_{t\to\infty} f(t)/\log(t) = 1$, we will allow f to have faster growth rates, including poly-logarithmic and polynomial rates.

Algorithm 1 (from Algorithm 2 of Cappé et al. (2013))

input: divergence $d_P : \mathcal{I}_P \times \mathcal{I}_P \to [0,\infty)$, increasing "exploration" function $f : (1,\infty) \to (0,\infty)$

initialize: Play each arm $1,\ldots,K$ once

for $t \geq K$ **do**

Play the arm (with ties broken arbitrarily):

$$A(t+1) = \arg\max_{i\in[K]} \sup\left\{ z \in \mathcal{I}_P : d_P(\widehat{\mu}_i(t), z) \leq \frac{f(t)}{N_i(t)} \right\}$$

end for

Below, we have Proposition 7, the main result of the section. In (34), we characterize the regret tail for Algorithm 1 under different choices of the exploration function f. We require some regularity conditions on f in order to be able to obtain the exact asymptotic limit in (34). We will use the notions of regularly varying and slowly varying functions (see, for example, Embrechts et al. (1997)). A function $h : \mathbb{R} \to \mathbb{R}$ is regularly varying if $h(x) = x^a l(x)$ for some $a \in \mathbb{R}$ (which is allowed to be zero), with l being a slowly varying function. And a function $l : \mathbb{R} \to \mathbb{R}$ is slowly varying if $\lim_{x\to\infty} l(cx)/l(x) = 1$ for any $c > 0$. Slowly varying functions can be intuitively thought of as functions with growth (or decay) that is slower than any polynomial (or inverse polynomial) (for example, any function with logarithmic or poly-logarithmic growth/decay).

The regret tail characterization in (34) is accompanied by an asymptotic characterization of expected regret in (35). To upper bound expected regret for Algorithm 1, a non-asymptotic minimum growth rate is required for f. Here, we use the rate: $f(t) \geq \log(1 + t\log^2(t))$ (for sufficiently large t). This is motivated by the development of KL-UCB in Chapter 10 of Lattimore and Szepesvári (2020) (see their Algorithm 8), where the exploration function is chosen to be $t \mapsto \log(1+t\log^2(t))$. Other choices are possible, including $t \mapsto \log(t) + 3\log\log(t)$, as used by Cappé et al. (2013) in the original development of KL-UCB. Such non-asymptotic minimum growth rates for f imply the asymptotic rate: $\liminf_{t\to\infty} f(t)/\log(t) \geq 1$, which is used to obtain the regret tail characterization in (34).

To establish (34) in Proposition 7, we adapt the proofs of Theorems 1 and 2. To establish (35), the asymptotic upper bound for expected regret follows from standard techniques found in, for example, Chapter 10 of Lattimore and Szepesvári (2020). The matching asymptotic lower bound for expected regret can be deduced using a LLN for the regret of Algorithm 1 (which follows directly from the proof of Lemma 3 in Appendix EC.4), together with Markov's inequality. The complete proof details for Proposition 7 can be found in Appendix EC.5.

PROPOSITION 7. *Let π be Algorithm 1, with divergence d_P and exploration function f.*

(i) Let f satisfy $\liminf_{t\to\infty} f(t)/\log(t) \geq 1$ and $f(t) = o(t^\lambda)$ for some $\lambda \in (0,1)$, and also let f be regularly varying and strictly increasing. Then, for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ and the i-th-best arm $r(i)$,

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{f(x)} = -\sum_{j=1}^{i-1} \inf_{z\in\mathcal{I}_P\,:\, z<\mu_{r(i)}} \frac{d_P(z,\mu_{r(j)})}{d_P(z,\mu_{r(i)})} \leq -(i-1) \tag{34}$$

uniformly for $x \in [f^{1+\gamma}(T), (1-\gamma)T]$ for any $\gamma \in (0, (1/\lambda)-1)$ as $T\to\infty$.

(ii) Let f satisfy $f(t) \geq \log(1+t\log^2(t))$ for sufficiently large t, and also $f(t) = o(t)$. Then, for any environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K}) \in \mathcal{M}_P^K$ and any sub-optimal arm i,

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{f(T)} = \frac{1}{d_P(\mu_i, \mu_*(\nu))}. \tag{35}$$

From (34), we see there is an explicit correspondence between the rate of exploration and the resulting heaviness of the regret distribution tail. For KL-UCB, the exploration function f satisfies $\lim_{t\to\infty} f(t)/\log(t) = 1$. In Algorithm 1, if we increase the rate of exploration to $1+b$ times the nominal rate of KL-UCB, so that $\lim_{t\to\infty} f(t)/\log(t) = 1+b$ (with $b>0$), then we obtain a regret tail exponent $\leq -(1+b)$. (See Figure 4 in Section 7 for numerical illustrations with different values of $b>0$ when $\mathcal{M}_P$ is a Gaussian family.) If we further increase the rate of exploration so that $\liminf_{t\to\infty} f(t)/t^a > 0$ for some $a \in (0,1)$, then the regret tail will decay at an exponential rate, as indicated by (34).

In summary, we are able to achieve essentially any desired regret tail using Algorithm 1 by suitably adjusting the rate of exploration via the exploration function f. Moreover, we will show in Sections 6.2-6.3 that an increased rate of exploration/a lighter regret tail provides increased robustness to model mis-specification. Of course, as indicated by (35), the price to pay is that the expected regret will increase accordingly as we use f with larger growth rates.

REMARK 4. Together, (34)-(35) in Proposition 7 provide a setting in which the trade-off in Theorem 4 between the regret tail in (31) and expected regret in (33) is tight. Specifically, tightness

is guaranteed when the model is an exponential family ($\mathcal{M} = \mathcal{M}_P$), and f is a slowly varying and strictly increasing function satisfying a minimum growth rate such as: $f(t) \geq \log(1 + t\log^2(t))$ (used to upper bound the expected regret). In this setting, the exact limit in (35) indicates that the asymptotic lower bound in (33) is not vacuous.

When f is regularly varying, but not slowly varying, then the trade-off in Theorem 4 is no longer tight. In this case, $\lim_{T\to\infty} f(cT)/f(T) \neq 1$ for any positive $c \neq 1$. This causes the regret tail upper bound condition in (31) to be sensitive to any multiplicative scaling of T. Consequently, it is harder to establish a sharp correspondence between a condition such as (31) and a resulting lower bound for expected regret such as (33).

Nevertheless, Proposition 7 suggests that Theorem 4 exhibits the correct dependence on T in the trade-off. In particular, given (31), the left side of (33) has the correct normalization dependence on T, but the resulting constant on the right side may not be tight. To see this, consider an exponential family model $\mathcal{M}_P$ with base distribution P, and let $f(t) = t^a$ with $a \in (0,1)$ (which is regularly varying, but not slowly varying). Then, Proposition 7 indicates that (34) holds and also,

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{T^a} = \frac{1}{d_P(\mu_i, \mu_*(\nu))},$$

for each sub-optimal arm i in environment $\nu = (P^{\mu_1}, \ldots, P^{\mu_K})$. However, given (34), in this setting Theorem 4 only indicates that

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{T^a} \geq \frac{1}{K^a} \frac{1}{d_P(\mu_i, \mu_*(\nu))},$$

which is off by a factor of K^{-a}, where K is the number of arms.

Remark 5. While studying a related problem, Audibert et al. (2009) developed finite-time upper bounds on the tail of the regret distribution for the UCB1 algorithm (due to Auer et al. (2002)) in the bounded rewards setting, which are suggestive of the exploration-regret tail trade-off that we provide in (34). They study the case where the amount of UCB exploration is increased by a $1 + b$ multiplicative factor (similar to setting f such that $\lim_{t\to\infty} f(t)/\log(t) = 1 + b$ in Algorithm 1), so that the regret tail becomes lighter with more negative tail exponent $-c \cdot (1 + b)$ for some fixed $c > 0$. However, they do not develop matching lower bounds for the regret tail. Such lower bounds are a fundamental ingredient in establishing the nature of the trade-off.

Remark 6. Furthermore, we can ensure the same regret tail guarantees as in (34) for all environments in a class $\mathcal{M}_{P,0}^K$ that is larger than $\mathcal{M}_P^K$. Here, $\mathcal{M}_{P,0}$ is a general family of distributions, whose CGF's are dominated by those of $\mathcal{M}_P$:

$$\mathcal{M}_{P,0} = \{Q : \mu(Q) \in \mathcal{I}_P,\ \Lambda_Q(\theta) \leq \Lambda_{P^{\mu(Q)}}(\theta)\ \forall\, \theta \in \mathbb{R}\}. \tag{36}$$

(Recall that $\Lambda_{P^{\mu(Q)}}$ is the CGF of $P^{\mu(Q)}$, the distribution resulting from tilting P to have mean $\mu(Q)$, as in (1).) Below are two examples of $\mathcal{M}_P$ and $\mathcal{M}_{P,0}$.

EXAMPLE 7. Let $\mathcal{M}_P$ be the Gaussian family with variance σ^2. Then $\mathcal{M}_{P,0}$ is the family of all sub-Gaussian distributions with variance proxy σ^2. (We say Z is sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[e^{\theta(Z-\mathbb{E}[Z])}] \le e^{\sigma^2\theta^2/2}$ for all $\theta \in \mathbb{R}$.)

EXAMPLE 8. Let $\mathcal{M}_P$ be the Bernoulli family. Then $\mathcal{M}_{P,0}$ is the family of all distributions supported on a subset of $[0,1]$.

6.2. Robustness to Mis-specified Reward Distribution

For an $\mathcal{M}_P$-optimized algorithm, if the true reward distributions do not belong in $\mathcal{M}_P$, then the regret tails can be heavier than truncated Cauchy, resulting in expected regret that grows polynomially in the time horizon. As we saw in Section 4.2 via Proposition 4 and Corollary 2, one example of this is when the variance in the design of KL-UCB for Gaussian bandits is just slightly under-specified relative to the true variance.

To alleviate such issues, we can use Algorithm 1 with a suitably increased rate of exploration through the choice of f, which leads to a lighter regret tail in well-specified settings, as previously shown in Proposition 7. We will see in Corollary 5 below that this also provides protection against distributional mis-specification of the arm rewards. In particular, by increasing f to be $1+b$ times the nominal rate $\log(1+t\log^2(t))$ (minimum growth rate of $f(t)$) used in part (ii) of Proposition 7, we can preserve logarithmic expected regret for environments from an enlarged class $\mathcal{M}_{P,b}^K$, which is defined in (37) below. Moreover, from part (i) of Proposition 7, in the well-specified case that the environment is from $\mathcal{M}_P^K$, the regret tail will have an exponent $\le -(1+b)$.

The enlarged family of distributions is

$$\mathcal{M}_{P,b} = \{Q : \mu(Q) \in \mathcal{I}_P,\ \Lambda_Q(\theta) \le \Psi_{P,b}(\mu(Q),\theta)\ \forall\, \theta \in \mathbb{R}\}, \tag{37}$$

where for any distribution Q and $z \in \mathcal{I}_Q$, we define:

$$\Psi_{Q,b}(z,\theta) = \frac{\Lambda_{Q^z}((1+b)\theta)}{1+b}, \quad \theta \in \mathbb{R}. \tag{38}$$

Setting $b=0$ recovers $\mathcal{M}_{P,0}$ as in (36). Using Jensen's inequality and the definition in (38), it is straightforward to see that $\mathcal{M}_P \subsetneqq \mathcal{M}_{P,b}$ for $b>0$. Moreover, with $\mathcal{M}_{P,0}$ from (36) and any $b'>b>0$, we have $\mathcal{M}_{P,0} \subsetneqq \mathcal{M}_{P,b} \subsetneqq \mathcal{M}_{P,b'}$. An example of $\mathcal{M}_P$ and $\mathcal{M}_{P,b}$ is the following.

EXAMPLE 9. Let $\mathcal{M}_P$ be the Gaussian family with variance σ^2. Then $\mathcal{M}_{P,b}$ is the family of all sub-Gaussian distributions with variance proxy $\sigma^2(1+b)$. (Also, $\mathcal{M}_{P,0}$ is the family of all sub-Gaussian distributions with variance proxy σ^2, as we saw in Example 7.)

The proof of Corollary 5 follows from a straightforward adaptation of the proof of (35) in Proposition 7. The details are provided in Appendix EC.6.

COROLLARY 5. *Let π be Algorithm 1, with divergence d_P and exploration function f. Let f satisfy $f(t) \geq (1+b)\log(1+t\log^2(t))$ with $b \geq 0$ for sufficiently large t, and also $f(t) = o(t)$. Then, for any environment $\nu = (Q_1, \ldots, Q_K) \in \mathcal{M}_{P,b}^K$ and each sub-optimal arm i,*

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{f(T)} = \frac{1}{d_P(\mu(Q_i), \mu_*(\nu))}. \tag{39}$$

6.3. Robustness to Mis-specified Reward Dependence Structure

In this section, we consider arm rewards taking values in a finite set $S \subset \mathbb{R}$. Let P be a distribution on S. Even if the marginal distributions of arm rewards belong in the exponential family $\mathcal{M}_P$, the serial dependence structure of rewards could be mis-specified, which can result in regret tails that are heavier than truncated Cauchy, and expected regret that grows polynomially in the time horizon. We saw an example of this in Section 4.4 via Proposition 5, particularly via Example 5.

To alleviate such issues, we can use Algorithm 1 with a suitably increased rate of exploration through the choice of f, which leads to a lighter regret tail in well-specified settings, as previously shown in Proposition 7. We will see in Corollary 6 below that this also provides protection against Markovian departures from independence of the arm rewards. In particular, by increasing f to be $1+b$ times the nominal rate $\log(1+t\log^2(t))$ (minimum growth rate of $f(t)$) used in part (ii) of Proposition 7, we can preserve logarithmic expected regret when the arm rewards evolve as Markov chains with transition matrices from a set $\widetilde{\mathcal{M}}_{P,b}$, which is defined in (40) below. Moreover, as in the previous section, in the well-specified case that the environment is from $\mathcal{M}_P^K$, the regret tail will have an exponent $\leq -(1+b)$.

Let $\mathcal{S}_{|S|}$ denote the set of $|S| \times |S|$ irreducible stochastic matrices satisfying the Doeblin Condition (as discussed in Section 4.4 in the context of Proposition 5). We define

$$\widetilde{\mathcal{M}}_{P,b} = \left\{ H \in \mathcal{S}_{|S|} : \phi_H(\theta) \leq \Psi_{P,b}(\phi'_H(0), \theta) \ \forall\, \theta \in \mathbb{R} \right\}, \tag{40}$$

and we recall that $\phi_H(\theta)$ is the logarithm of the Perron-Frobenius eigenvalue of the tilted version (as in (26)) of transition matrix H, and $\phi'_H(0)$ is the equilibrium mean of a chain with transition matrix H. Of course, the exponential family $\mathcal{M}_P$ is equivalent to a strict subset of the collection of transition matrices with identical rows in $\widetilde{\mathcal{M}}_{P,b}$, for any $b > 0$. Also, for any $b' > b > 0$, $\widetilde{\mathcal{M}}_{P,b} \subsetneq \widetilde{\mathcal{M}}_{P,b'}$. In Example 10, which is given after Corollary 6, we examine the degree to which $\widetilde{\mathcal{M}}_{P,b}$ is "larger" than $\mathcal{M}_P$ when $S = \{0,1\}$ and $\mathcal{M}_P$ is the Bernoulli family.

Like for Corollary 5, the proof of Corollary 6 also follows from a straightforward adaptation of the proof of (35) in Proposition 7. The details are provided in Appendix EC.6.

COROLLARY 6. *Let π be Algorithm 1, with divergence d_P and exploration function f. Let f satisfy $f(t) \geq (1+b)\log(1+t\log^2(t))$ with $b \geq 0$ for sufficiently large t, and also $f(t) = o(t)$. For the K-armed environment ν, suppose arm i yields rewards that evolve according to a Markov chain with transition matrix $H_i \in \widetilde{\mathcal{M}}_{P,b}$. Then, for any sub-optimal arm i,*

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{f(T)} = \frac{1}{d_P(\phi'_{H_i}(0), \phi'_{H_{r(1)}}(0))}. \tag{41}$$

EXAMPLE 10. Let the state space $S = \{0,1\}$, and let $\mathcal{M}_P$ be the Bernoulli family of distributions. Consider transition matrices on S of the form:

$$H = \begin{bmatrix} 1-q & q \\ 1-q' & q' \end{bmatrix}. \tag{42}$$

The more positive the difference $q'-q$, the more positive the autocorrelation between the rewards. In Table 2 below, for different values of $b > 0$, we examine how positive the difference $q'-q$ can be in order for H to still belong in $\widetilde{\mathcal{M}}_{P,b}$, and thus for Corollary 6 to be applicable. As the targeted regret tail exponent $-(1+b)$ is made more negative, the algorithm can withstand more positive autocorrelation between the rewards and still maintain logarithmic expected regret.

$-(1+b)$	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
max allowed $q'-q$	0.18	0.36	0.49	0.59	0.65	0.70	0.74	0.77	0.80	0.82

Table 2 For particular $-(1+b)$ values (upper bound on the regret tail exponent), and for the restriction $q, q' \in [0.05, 0.95]$, we give the maximum allowed difference $q'-q$ that ensures the transition matrix H in (42) belongs in $\widetilde{\mathcal{M}}_{P,b}$, as defined in (40) (and (38)).

REMARK 7. For general reward processes satisfying Assumptions 1-2, e.g., general Markov processes, there are no finite-sample concentration bounds. So there does not seem to be a universal way to obtain an upper bound on the regret tail. For such reward processes, there also does not seem to be a universal way to obtain upper bounds on expected regret such as in Corollary 6, and thus there are no provable robustness benefits for our procedure to lighten the regret tail. Nevertheless, our simulations in Figure 5 in Section 7 suggest that we can still ensure the regret tail is lighter to a desired level using our procedure. (So, the lower bound in Theorem 3 seems to be tight in greater generality than what we are able to provably show.)

7. Numerical Experiments

In this section, we use numerical experiments to verify that our asymptotic approximations for the regret distribution tail hold over finite time horizons. For each experiment, we perform (statistically) independent simulation runs of the bandit system, and we compute empirical probabilities for the regret tail.

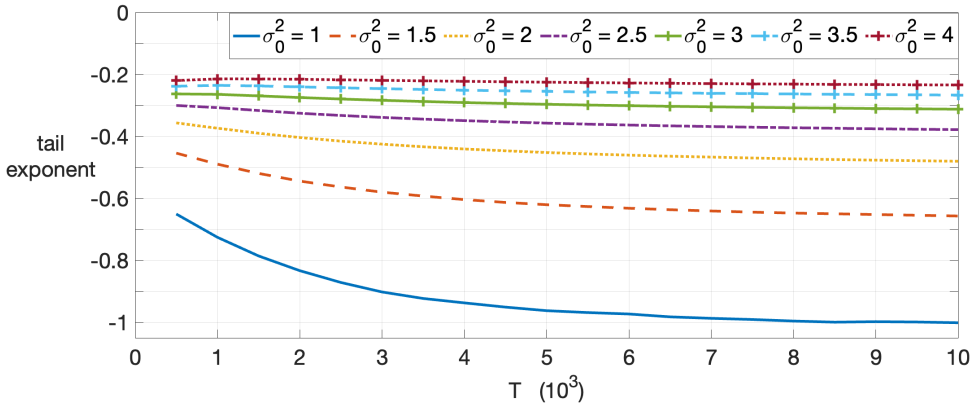

Figure 1 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) \geq 0.8T)/\log(T)$ vs T. Environment $\nu = (N(0.1, \sigma_0^2), N(0, \sigma_0^2))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards. The curves correspond to the cases $\sigma_0^2 = 1, 1.5, \ldots, 4$, as indicated by the legend. The curves asymptote to $-1/\sigma_0^2$ in each case, which agrees with (20) in Corollary 2. To generate each curve, 2×10^6 simulation runs were used.

In Figure 1, we examine the validity of Theorem 1 and Corollary 2. For all curves but the dark blue one, the variance of the Gaussian KL-UCB algorithm is set smaller than that of the actual Gaussian reward distributions. In Figure 2, we examine the validity of Corollary 3. For all curves but the dark blue one, the Gaussian KL-UCB algorithm does not take into account the AR(1) serial dependence structure of the rewards, even though the algorithm is perfectly matched to the marginal distributions of the rewards. In both Figures 1 and 2, the regret tail probabilities in mis-specified cases correspond to regret distribution tails that are heavier than truncated Cauchy.

In Figure 3, we verify that when the arms are iid Bernoulli, KL-UCB produces regret distribution tails which are strictly lighter than truncated Cauchy, as predicted by Theorem 2.

In Figure 4, we demonstrate the trade-off established in (34) of Proposition 7 between the amount of UCB exploration and the resulting exponent of the regret distribution tail, with $f(t) = (1+b)\log(t)$ and $b \geq 0$ in Algorithm 1.

In Figure 5, we demonstrate that the poor regret tail properties resulting from mis-specification of the serial dependence structure of the rewards can be overcome by aiming for a lighter regret tail using Algorithm 1. Here, we use the same AR(1) setup that is illustrated in Figure 2. As discussed in the first paragraph of Remark 7, here we do not have upper bounds on regret tail probabilities (only lower bounds in (25)), and thus there are no provable robustness guarantees. However, we show empirically in Figure 5 that aiming for a lighter regret tail still provides robustness to mis-specification in this setting. The $\frac{1+\beta_0}{1-\beta_0}$ factor in Figure 5 is taken from the lower bound in (25), which we essentially confirm to be tight here.

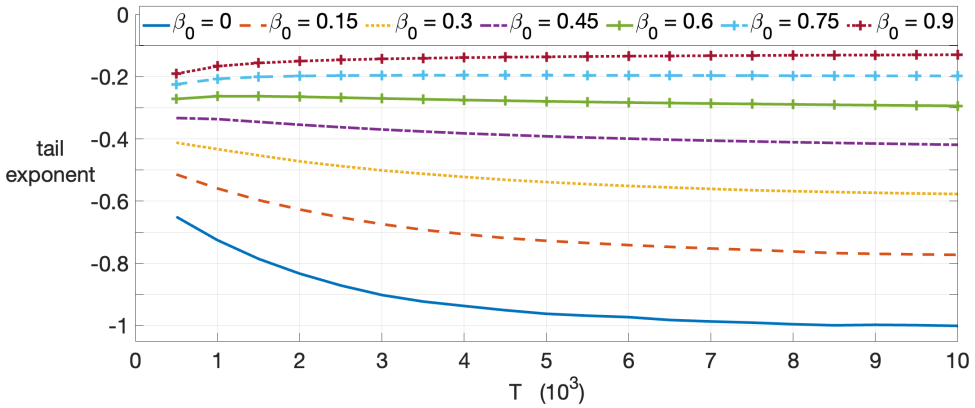

Figure 2 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) \geq 0.8T)/\log(T)$ vs T. Environment ν consists of two Gaussian AR(1) processes with common AR coefficient β_0, and equilibrium distributions $(N(0.1,1), N(0,1))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards. The curves correspond to the cases $\beta_0 = 0, 0.15, \ldots, 0.9$, as indicated by the legend. The curves approximately asymptote to $-(1-\beta_0)/(1+\beta_0)$, which agrees with the lower bound in Corollary 3 and (25). To generate each curve, 2×10^6 simulation runs were used.

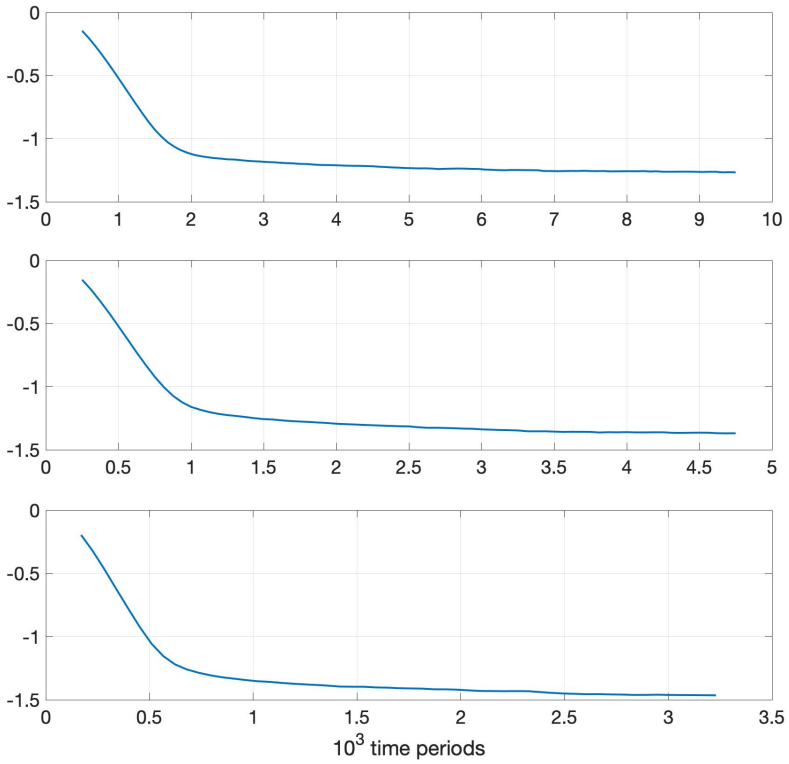

Figure 3 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$ (with time horizon T fixed). Environment $\nu = (\text{Ber}(q), \text{Ber}(0.4))$. Algorithm π is KL-UCB for iid Bernoulli rewards. Top: $q = 0.475$, $T = 10^4$; Middle: $q = 0.5$, $T = 5 \times 10^3$; Bottom: $q = 0.525$, $T = 3.4 \times 10^3$. Each curve asymptotes to $\lim_{z \downarrow 0} d_P(z,q)/d_P(z,0.4)$ (with values -1.26 (top), -1.36 (middle), -1.46 (bottom)), as specified by Theorem 2 and (10). To generate each curve, 8×10^6 simulation runs were used.

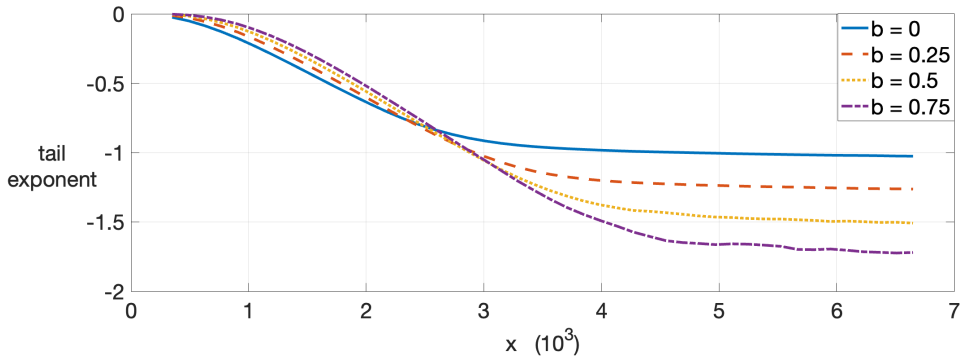

Figure 4 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$, with fixed time horizon $T = 7 \times 10^3$. Environment $\nu = (N(0.1,1), N(0,1))$. π is Algorithm 1 with KL divergence d_P between unit-variance Gaussian distributions, and $f(t) = (1+b)\log(t)$ (to aim for a regret tail exponent of $-(1+b)$). The curves correspond to the cases $b = 0, 0.25, 0.5, 0.75$, as indicated by the legend. As predicted by (34) in Proposition 7, the curves asymptote to -1, -1.25, -1.5, -1.75. To generate each curve, 4×10^7 simulation runs were used.

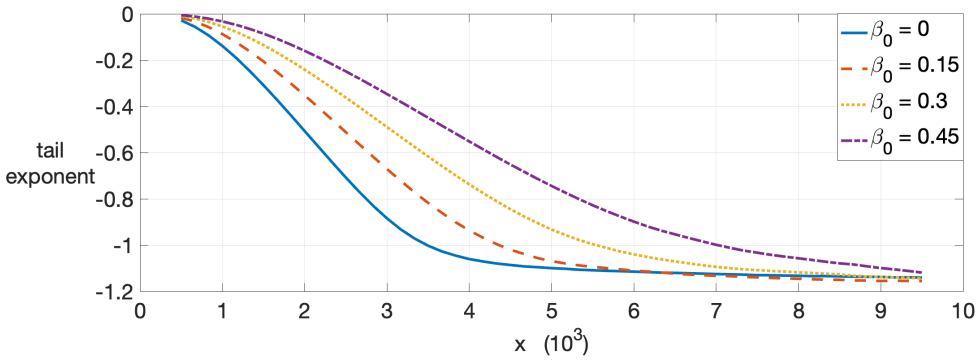

Figure 5 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$, with fixed time horizon $T = 10^4$. Environment ν consists of two Gaussian AR(1) processes with common AR coefficient β_0, and equilibrium distributions $(N(0.1,1), N(0,1))$. π is Algorithm 1 with KL divergence d_P between unit-variance Gaussian distributions, and $f(t) = (1+b)\log(t)$ with $1 + b = 1.1 \cdot \frac{1+\beta_0}{1-\beta_0}$ (to aim for a regret tail exponent of ≈ -1.1 in each case of β_0). The curves correspond to the cases $\beta_0 = 0, 0.15, 0.3, 0.45$, as indicated by the legend. All curves asymptote to (slightly less than) -1.1, as desired. To generate each curve, 4×10^7 simulation runs were used.

8. Acknowledgments

We thank the Area Editor, Ramandeep Randhawa, the Associate Editor, and two Reviewers for their valuable suggestions, which significantly improved the paper. We also thank Achal Bassamboo and Nalin Shani for a close reading of some of the technical arguments, which led to more streamlined arguments. We are indebted to Professor Tze Leung Lai, who sadly passed away in May 2023, for paving the way in numerous fundamental areas of research, extending well beyond

the multi-armed bandit problem studied here. Peter Glynn is also grateful to have had Tze Leung as a colleague and friend for many years.

Appendix A: Proof of Theorem 1

Proof of Theorem 1. Without loss of generality, suppose that $\mu_1 > \mu_2 > \cdots > \mu_K$ (i.e., $r(i) = i$ for all $i \in [K]$) in the environment $\nu = (P^{\mu_1}, P^{\mu_2}, \ldots, P^{\mu_K})$. We first show (7) and (8) for the second-best arm $i = 2$. Consider the alternative environment $\widetilde{\nu} = (P^{\widetilde{\mu}_1}, P^{\mu_2}, \ldots, P^{\mu_K})$, where $\widetilde{\mu}_1 < \mu_2$, and $\mu_2, \ldots, \mu_K$ are the same mean values from the original environment ν. (Arm 2 is the best arm in $\widetilde{\nu}$. Later in the proof, we will consider different values for $\widetilde{\mu}_1$, subject to $\widetilde{\mu}_1 < \mu_2$ and $\widetilde{\mu}_1 \in \mathcal{I}_P$.) Let $\delta > 0$, and define the events:

$$\mathcal{A}_T = \left\{ \left| \frac{N_1(T)}{\log(T)} - \frac{1}{d_P(\widetilde{\mu}_1, \mu_2)} \right| \le \delta \right\} \cap \left\{ \left| \frac{N_j(T)}{\log(T)} - \frac{1}{d_P(\mu_j, \mu_2)} \right| \le \delta, \, \forall \, j \ge 3 \right\}$$

$$\mathcal{B}_T = \left\{ \left| \frac{1}{N_1(T)} \sum_{t=1}^{N_1(T)} \log \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t)) + d_P(\widetilde{\mu}_1, \mu_1) \right| \le \delta \right\}.$$

By a change of measure from ν to $\widetilde{\nu}$,

$$\mathbb{P}_{\nu\pi}(N_2(T) > (1-\gamma)T) = \mathbb{E}_{\widetilde{\nu}\pi} \left[\mathbb{I}\left(N_2(T) > (1-\gamma)T\right) \prod_{t=1}^{N_1(T)} \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t)) \right] \tag{43}$$

$$\ge \mathbb{E}_{\widetilde{\nu}\pi} \left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(\frac{1}{N_1(T)} \sum_{t=1}^{N_1(T)} \log \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t)) \cdot N_1(T) \right) \right] \tag{44}$$

$$\ge \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T, \mathcal{B}_T) \cdot \exp\left(-\left(d_P(\widetilde{\mu}_1, \mu_1) + \delta\right) \left(\frac{1}{d_P(\widetilde{\mu}_1, \mu_2)} + \delta \right) \log(T) \right). \tag{45}$$

where (44) follows from $\{N_2(T) > (1-\gamma)T\} \supset \mathcal{A}_T$ for sufficiently large T, and (45) follows from lower bounds using $\mathcal{A}_T$ and $\mathcal{B}_T$. From (45), taking logs and dividing by $\log(T)$,

$$\frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > (1-\gamma)T)}{\log(T)} \ge \frac{\log \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T, \mathcal{B}_T)}{\log(T)} - \left(d_P(\widetilde{\mu}_1, \mu_1) + \delta\right) \left(\frac{1}{d_P(\widetilde{\mu}_1, \mu_2)} + \delta \right). \tag{46}$$

Using Lemma 2 together with the WLLN for sample means, we have $\lim_{T\to\infty} \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T, \mathcal{B}_T) = 1$. So the first term on the right side of (46) is negligible as $T \to \infty$, and upon sending $\delta \downarrow 0$ and optimizing with respect to $\widetilde{\mu}_1$, we have

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > (1-\gamma)T)}{\log(T)} \ge - \inf_{\widetilde{\mu}_1 \in \mathcal{I}_P \,:\, \widetilde{\mu}_1 < \mu_2} \frac{d_P(\widetilde{\mu}_1, \mu_1)}{d_P(\widetilde{\mu}_1, \mu_2)}. \tag{47}$$

Then, the conclusion (7) with the infimum over $B_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$ follows from part (i) of Lemma 5 (provided after the current proof) with the choice $g(t) = \log^{1+\gamma}(t)$.

We now establish (8). Let $\gamma \in (0,1)$. In the context of Lemma 5, take $g(t) = t^\gamma$. Because P is discrimination equivalent, the right side of (47) is equal to -1, which establishes part (i) of Lemma 5. Also, since π is $\mathcal{M}_P$-optimized, using Markov's inequality,

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > T^\gamma)}{\log(T^\gamma)} \le -1,$$

which establishes part (ii) of Lemma 5. Then, the desired uniform convergence in (8) for $x \in [T^\gamma, (1-\gamma)T]$ follows from part (iii) of Lemma 5.

We now show (7) for any sub-optimal arm $i \geq 3$. Consider the alternative environment $\widetilde{\nu} = (P^{\widetilde{\mu}_1}, \ldots, P^{\widetilde{\mu}_{i-1}}, P^{\mu_i}, \ldots, P^{\mu_K})$, where $\widetilde{\mu}_j < \mu_i$ for all $j \leq i-1$, and $\mu_i, \ldots, \mu_K$ are the same mean values from the original environment ν. (Arm i is now the best arm in $\widetilde{\nu}$.) We change the events $\mathcal{A}_T$ and $\mathcal{B}_T$ to:

$$\mathcal{A}_T = \left\{ \left| \frac{N_j(T)}{\log(T)} - \frac{1}{d_P(\widetilde{\mu}_j, \mu_i)} \right| \leq \delta, \, \forall \, j \leq i-1 \right\} \cap \left\{ \left| \frac{N_j(T)}{\log(T)} - \frac{1}{d_P(\mu_j, \mu_i)} \right| \leq \delta, \, \forall \, j \geq i+1 \right\}$$

$$\mathcal{B}_T = \left\{ \left| \frac{1}{N_j(T)} \sum_{t=1}^{N_j(T)} \log \frac{dP^{\mu_j}}{dP^{\widetilde{\mu}_j}}(X_j(t)) + d_P(\widetilde{\mu}_j, \mu_j) \right| \leq \delta, \, \forall \, j \leq i-1 \right\}.$$

To obtain (7) for sub-optimal arm $i \geq 3$, we can then run through arguments analogous to those in (43)-(47). Here, the change of measure from ν to $\widetilde{\nu}$ involves the product of $i-1$ likelihood ratios corresponding to the arms $1, \ldots, i-1$. Each of the parameter values $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_{i-1}$ can be optimized separately (subject to $\widetilde{\mu}_j < \mu_i$ and $\widetilde{\mu}_j \in \mathcal{I}_P$ for all $j \leq i-1$) to yield the desired conclusion. □

We now introduce Lemma 5, which provides a unified way to establish *uniform convergence* for the regret tail characterizations developed throughout the paper.

LEMMA 5. *Let ν be any bandit environment, and let i be a sub-optimal arm in ν.*

(i) Suppose for some $c_i(\nu) > 0$ and some $\gamma \in (0,1)$,

$$\liminf_{T\to\infty} \frac{\log P_{\nu\pi}(N_i(T) > (1-\gamma)T)}{\log(T)} \geq -c_i(\nu). \tag{48}$$

Then,

$$\liminf_{T\to\infty} \inf_{x \in B_\gamma(T)} \frac{\log P_{\nu\pi}(N_i(T) > x)}{\log(x)} \geq -c_i(\nu), \tag{49}$$

with $B_\gamma(T) = [g(T), (1-\gamma)T]$, and any strictly increasing function $g : (1,\infty) \to (0,\infty)$ such that $\lim_{t\to\infty} g(t)/\log(t) = \infty$ and $g(t) = o(t)$.

(ii) Suppose,

$$\limsup_{T\to\infty} \frac{\log P_{\nu\pi}(N_i(T) > g(T))}{\log(g(T))} \leq -c_i(\nu). \tag{50}$$

Then,

$$\limsup_{T\to\infty} \sup_{x \in B_\gamma(T)} \frac{\log P_{\nu\pi}(N_i(T) > x)}{\log(x)} \leq -c_i(\nu). \tag{51}$$

(iii) If both (48) and (50) hold, then together (49) and (51) yield:

$$\lim_{T\to\infty} \frac{\log P_{\nu\pi}(N_i(T) > x)}{\log(x)} = -c_i(\nu)$$

uniformly for $x \in B_\gamma(T)$ as $T \to \infty$.

Proof of Lemma 5.

Part (i)

Since $t \mapsto N_i(t)$ is non-decreasing with $\mathbb{P}_{\nu\pi}$-probability one, we have for sufficiently large T and all $x \in B_\gamma(T) = [g(T), (1-\gamma)T]$,

$$\mathbb{P}_{\nu\pi}(N_i(T) > x) \geq \mathbb{P}_{\nu\pi}(N_i(\lceil x/(1-\gamma)\rceil) > x).$$

So, for any $\epsilon > 0$, we have for sufficiently large T,

$$\frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > x)}{\log(x)} \geq \frac{\log \mathbb{P}_{\nu\pi}(N_i(\lceil x/(1-\gamma)\rceil) > x)}{\log(x)} \tag{52}$$

$$\geq -c_i(\nu)(1+\epsilon), \tag{53}$$

uniformly for all $x \in B_\gamma(T)$, where (53) follows from the convergence result in (48). Then, (49) is established by taking the infimum over $x \in B_\gamma(T)$ on the left side of (52), sending $T \to \infty$, and then $\epsilon \downarrow 0$.

Part (ii)

The function g is strictly increasing, so it has an inverse g^{-1} (defined on the range of g), which is also strictly increasing. Since $t \mapsto N_i(t)$ is non-decreasing with $\mathbb{P}_{\nu\pi}$-probability one, we have for sufficiently large T and all $x \in B_\gamma(T) = [g(T), (1-\gamma)T]$,

$$\mathbb{P}_{\nu\pi}(N_i(\lfloor g^{-1}(x)\rfloor) > x) \geq \mathbb{P}_{\nu\pi}(N_i(T) > x).$$

So, for any $\epsilon \in (0,1)$, we have for sufficiently large T,

$$-c_i(\nu)(1-\epsilon) \geq \frac{\log \mathbb{P}_{\nu\pi}(N_i(\lfloor g^{-1}(x)\rfloor) > x)}{\log(x)} \tag{54}$$

$$\geq \frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > x)}{\log(x)}, \tag{55}$$

uniformly for all $x \in B_\gamma(T)$, where (54) follows from the convergence result in (50). Then, (51) is established by taking the supremum over $x \in B_\gamma(T)$ on the right side of (55), sending $T \to \infty$, and then $\epsilon \downarrow 0$. □

Appendix B: Proof of Theorem 2

Define for arm i the UCB index at time t, given that arm i has been played n times:

$$\widetilde{U}_i(n,t) = \sup\left\{ z \in \mathcal{I}_P : d_P(\widehat{\mu}_i(\tau_i(n)), z) \leq \frac{f(t)}{n} \right\},$$

where $\tau_i(n)$ is the time of the n-th play of arm i. The "exploration" function $f(t)$ is a design choice. For KL-UCB (as introduced in Algorithm 2 of Cappé et al. (2013)), choices include $f(t) = \log(t)$

(as in Section 7 of Cappé et al. (2013)), $f(t) = \log(t) + 3\log\log(t)$ (as in Theorem 1 of Cappé et al. (2013)), and $f(t) = \log(1 + t\log^2(t))$ (as in Theorem 10.6 of Lattimore and Szepesvári (2020)). We will leave the particular form for $f(t)$ unspecified in developing the upper bound part of this proof. We will actually establish a more general upper bound part. Specifically, we consider any increasing function $f : (1,\infty) \to (0,\infty)$ such that $\liminf_{t\to\infty} f(t)/\log(t) \geq 1$ and $f(t) = o(t^\lambda)$ for some $\lambda \in (0,1)$.

Proof of Theorem 2. Without loss of generality, suppose that $\mu_1 > \mu_2 > \cdots > \mu_K$ (i.e., $r(i) = i$ for all $i \in [K]$) in the environment $\nu = (P^{\mu_1}, P^{\mu_2}, \ldots, P^{\mu_K})$.

Upper Bound

We first consider the sub-optimal arm $i = 2$. Let $x_T = \lfloor f^{1+\gamma}(T) \rfloor$ with any fixed $\gamma \in (0, (1/\lambda) - 1)$ (with $\lambda \in (0,1)$ as specified above). Also, let $\delta \in (0, \mu_1 - \mu_2)$. We have the following bounds:

$$\begin{aligned}
\mathbb{P}_{\nu\pi}(N_2(T) > x_T) &\leq \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (\tau_2(x_T), T] \ \text{ s.t. } \ \widetilde{U}_1(N_1(t-1), t-1) \leq \widetilde{U}_2(N_2(t-1), t-1)\Big) && (56)\\
&\leq \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \ \text{ s.t. } \ \widetilde{U}_1(N_1(t-1), x_T) \leq \widetilde{U}_2(x_T, T)\Big)\\
&\leq \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \ \text{ s.t. } \ \widetilde{U}_1(N_1(t-1), x_T) \leq \mu_2 + \delta\Big) && (57)\\
&\quad + \mathbb{P}_{\nu\pi}\Big(\widetilde{U}_2(x_T, T) > \mu_2 + \delta\Big). && (58)
\end{aligned}$$

Note that (56) holds because $N_2(T) > x_T$ is the event of interest, and so after the x_T-th play of arm 2 at time $\tau_2(x_T)$, there must be at least one more time period in which arm 2 is played. In that period, the UCB index of arm 2 must be greater than that of arm 1 (and that of all other arms).

For the term in (57), we have

$$\begin{aligned}
(57) &\leq \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\Big(\widetilde{U}_1(m, x_T) \leq \mu_2 + \delta\Big) && (59)\\
&= \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\left(d_P(\widehat{\mu}_1(\tau_1(m)), \mu_2 + \delta) \geq \frac{f(x_T)}{m},\ \ \widehat{\mu}_1(\tau_1(m)) < \mu_2 + \delta\right)\\
&= \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\left(\frac{1}{m}\sum_{l=1}^{m} X_1(l) \leq y_m^*\right)\\
&\leq \sum_{m=1}^{\infty} \exp\left(-m \cdot d_P(y_m^*, \mu_1)\right), && (60)
\end{aligned}$$

where for each m, y_m^* is the unique solution to $d_P(y_m^*, \mu_2 + \delta) = f(x_T)/m$ and $y_m^* < \mu_2 + \delta$, and we have used a Chernoff bound in (60). We define

$$s_T = \frac{2f(x_T)}{d_P(\mu_2 + \delta, \mu_1)} \cdot \inf_{z < \mu_2 + \delta} \frac{d_P(z, \mu_1)}{d_P(z, \mu_2 + \delta)},$$

and so for $m \geq s_T$,

$$\frac{d_P(\mu_2+\delta,\mu_1)}{2} \geq \frac{f(x_T)}{m} \cdot \inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)}.$$

Since $y_m^* < \mu_2+\delta$, we have $d_P(y_m^*,\mu_1) \geq d_P(\mu_2+\delta,\mu_1)$, and so for $m \geq s_T$,

$$d_P(y_m^*,\mu_1) \geq \frac{f(x_T)}{m} \cdot \inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} + \frac{d_P(\mu_2+\delta,\mu_1)}{2}. \tag{61}$$

Splitting the sum in (60) into two pieces at s_T, we have

$$\begin{aligned}
(60) &= \sum_{m=1}^{\lfloor s_T \rfloor} \exp\left(-m \cdot d_P(y_m^*,\mu_2+\delta) \cdot \frac{d_P(y_m^*,\mu_1)}{d_P(y_m^*,\mu_2+\delta)}\right) \\
&\quad + \sum_{m=\lfloor s_T \rfloor+1}^{\infty} \exp\left(-m \cdot d_P(y_m^*,\mu_1)\right) \\
&\leq \sum_{m=1}^{\lfloor s_T \rfloor} \exp\left(-m \cdot \frac{f(x_T)}{m} \cdot \inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)}\right) && (62) \\
&\quad + \sum_{m=\lfloor s_T \rfloor+1}^{\infty} \exp\left(-m \cdot \left(\frac{f(x_T)}{m} \cdot \inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} + \frac{d_P(\mu_2+\delta,\mu_1)}{2}\right)\right) && (63) \\
&= \exp\left(-f(x_T) \cdot \inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)}\right) \left(\lfloor s_T \rfloor + \sum_{m=\lfloor s_T \rfloor+1}^{\infty} \exp\left(-m \cdot \frac{d_P(\mu_2+\delta,\mu_1)}{2}\right)\right). && (64)
\end{aligned}$$

In (62), we use the fact that $d_P(y_m^*,\mu_2+\delta) = f(x_T)/m$. In (63), we use (61) (for $m \geq s_T$).

For the term in (58), we have for sufficiently large T,

$$\left|\widetilde{U}_2(x_T,T) - \widehat{\mu}_2(\tau_2(x_T))\right| < \frac{\delta}{2}.$$

So, for sufficiently large T,

$$\begin{aligned}
(58) &\leq \mathbb{P}_{\nu\pi}\left(\frac{1}{x_T}\sum_{l=1}^{x_T} X_2(l) > \mu_2 + \frac{\delta}{2}\right) \\
&\leq \exp\Big(-x_T \cdot d_P(\mu_2+\delta/2,\mu_2)\Big), && (65)
\end{aligned}$$

where (65) follows from a Chernoff bound.

Using (57) and (64) together with (58) and (65), we have

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}\left(N_2(T) > x_T\right)}{f(x_T)} \leq -\inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)}. \tag{66}$$

From the argument included separately in Appendix EC.2,

$$\lim_{\delta\downarrow 0} \inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} = \inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2)}. \tag{67}$$

Thus, we have established that for sub-optimal arm $i=2$,

$$\limsup_{T\to\infty}\frac{\log \mathbb{P}_{\nu\pi}\left(N_2(T)>x_T\right)}{f(x_T)}\leq -\inf_{z<\mu_2}\frac{d_P(z,\mu_1)}{d_P(z,\mu_2)}.$$

We now consider any sub-optimal arm $i\geq 3$. Let $\delta\in(0,\mu_{i-1}-\mu_i)$. In parallel to (57) and (58), we have

$$\mathbb{P}_{\nu\pi}\left(N_i(T)>x_T\right)\leq \mathbb{P}_{\nu\pi}\Big(\exists\, t\in(x_T,T] \text{ s.t. } \max_{1\leq j\leq i-1}\widetilde{U}_j(N_j(t-1),x_T)\leq \mu_i+\delta\Big) \tag{68}$$

$$+\,\mathbb{P}_{\nu\pi}\Big(\widetilde{U}_i(x_T,T)>\mu_i+\delta\Big). \tag{69}$$

We can bound (68) via:

$$\begin{aligned}(68)&\leq \mathbb{P}_{\nu\pi}\Big(\forall\, 1\leq j\leq i-1,\ \ \exists\, m_j\in\mathbb{Z}_+ \text{ s.t. } \widetilde{U}_j(m_j,x_T)\leq \mu_i+\delta\Big)\\ &\leq \prod_{j=1}^{i-1}\sum_{m=1}^{\infty}\mathbb{P}_{\nu\pi}\Big(\widetilde{U}_j(m,x_T)\leq \mu_i+\delta\Big),\end{aligned} \tag{70}$$

where (70) follows from the independence of the rewards from different arms. We can then upper bound each term of the product in (70) in the same way that we upper bounded (59). We can upper bound (69) in the same way that we upper bounded (58), and thus show that (69) is asymptotically negligible. Following the rest of the argument above (which was for the case $i=2$), we obtain for any sub-optimal arm $i\geq 3$:

$$\limsup_{T\to\infty}\frac{\log \mathbb{P}_{\nu\pi}\left(N_i(T)>x_T\right)}{f(x_T)}\leq -\sum_{j=1}^{i-1}\inf_{z<\mu_i}\frac{d_P(z,\mu_j)}{d_P(z,\mu_i)},$$

where the sum on the right side results from taking log of the product in (70).

Lower Bound and Final Result

Consider any sub-optimal arm $i\geq 2$. In the context of Lemma 5, take $g(t)=\log^{1+\gamma}(t)$ with any $\gamma>0$. From proof of Theorem 1, we have the lower bound:

$$\liminf_{T\to\infty}\frac{\log \mathbb{P}_{\nu\pi}\left(N_i(T)>(1-\gamma)T\right)}{\log(T)}\geq -\sum_{j=1}^{i-1}\inf_{z<\mu_i}\frac{d_P(z,\mu_j)}{d_P(z,\mu_i)}, \tag{71}$$

which establishes part (i) of Lemma 5.

In the above proof of the upper bound part, choose increasing $f:(1,\infty)\to(0,\infty)$ to satisfy $\lim_{t\to\infty} f(t)/\log(t)=1$, with $x_T=\lfloor g(T)\rfloor$. Then, the above proof of the upper bound part yields for any sub-optimal arm $i\geq 2$:

$$\limsup_{T\to\infty}\frac{\log \mathbb{P}_{\nu\pi}\left(N_i(T)>g(T)\right)}{\log(g(T))}\leq -\sum_{j=1}^{i-1}\inf_{z<\mu_i}\frac{d_P(z,\mu_j)}{d_P(z,\mu_i)}, \tag{72}$$

which establishes part (ii) of Lemma 5.

The desired uniform convergence result in (18) then follows from part (iii) of Lemma 5.

□

Appendix C: Proof of Theorem 3

Proof of Theorem 3. Without loss of generality, suppose that the long-run average rewards (in the sense of (21)) for arms $1,2,\ldots,K$ within the environment ν satisfy $\overline{\Lambda}_1'(0) > \overline{\Lambda}_2'(0) > \cdots > \overline{\Lambda}_K'(0)$ (i.e., $r(i) = i$ for all $i \in [K]$). Consider any sub-optimal arm $i \geq 2$. Let $\widetilde{\nu}$ be an alternative environment where the reward distribution structure remains the same for arms $i, i+1, \ldots, K$. However, for arms $j \leq i-1$ in the environment $\widetilde{\nu}$, let the distribution of $\sum_{t=1}^n X_j(t)$ for each $n \geq 1$ be

$$Q_j^n(dx;\theta_j) = \exp\left(\theta_j \cdot x - n\overline{\Lambda}_j^n(\theta_j)\right) Q_j^n(dx),$$

where $Q_j^n(dx)$ is the original distribution for $\sum_{t=1}^n X_j(t)$ in the environment ν. Moreover, let $\theta_j \in \overline{\Theta}_j$ such that $\overline{\Lambda}_j'(\theta_j) < \overline{\Lambda}_i'(0)$ (note that $\theta_j < 0$). (If this is not possible, then the infimum on the right side of (23) is empty, and the lower bound is $-\infty$.) So, in the environment $\widetilde{\nu}$, arm i yields the greatest long-run average rewards compared to all other arms.

Let $\delta > 0$, and define the events:

$$\begin{aligned}
\mathcal{A}_T &= \left\{ \left| \frac{N_j(T)}{\log(T)} - \frac{1}{d_P(\overline{\Lambda}_j'(\theta_j), \overline{\Lambda}_i'(0))} \right| \leq \delta, \, \forall\, j \leq i-1 \right\} \\
&\quad \cap \left\{ \left| \frac{N_j(T)}{\log(T)} - \frac{1}{d_P(\overline{\Lambda}_j'(0), \overline{\Lambda}_i'(0))} \right| \leq \delta, \, \forall\, j \geq i+1 \right\} \\
\mathcal{B}_T &= \left\{ \left| \widehat{\mu}_j(T) - \overline{\Lambda}_j'(\theta_j)) \right| \leq \delta, \, \forall\, j \leq i-1 \right\}.
\end{aligned}$$

Following steps analogous to (43)-(45) in the proof of Theorem 1,

$$\begin{aligned}
&\mathbb{P}_{\nu\pi}(N_i(T) > (1-\gamma)T) \\
&= \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(N_i(T) > (1-\gamma)T\right) \exp\left(\sum_{j=1}^{i-1} \left(-\theta_j \cdot \sum_{t=1}^{N_j(T)} X_j(t) + N_j(T) \cdot \overline{\Lambda}_j^{N_j(T)}(\theta_j) \right) \right) \right] && (73) \\
&\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(\sum_{j=1}^{i-1} \left(-\theta_j \cdot \widehat{\mu}_j(T) + \overline{\Lambda}_j^{N_j(T)}(\theta_j) \right) N_j(T) \right) \right] && (74) \\
&\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(\sum_{j=1}^{i-1} \left(-\theta_j \cdot \left(\overline{\Lambda}_j'(\theta_j) - \delta\right) + \overline{\Lambda}_j(\theta_j) - \delta \right) N_j(T) \right) \right] && (75) \\
&= \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(-\sum_{j=1}^{i-1} \left(\overline{\Lambda}_j^*(\overline{\Lambda}_j'(\theta_j)) + \delta(1-\theta_j) \right) N_j(T) \right) \right] && (76) \\
&\geq \mathbb{P}_{\widetilde{\nu}\pi}\left(\mathcal{A}_T, \mathcal{B}_T\right) \cdot \exp\left(-\sum_{j=1}^{i-1} \left(\overline{\Lambda}_j^*(\overline{\Lambda}_j'(\theta_j)) + \delta(1-\theta_j) \right) \left(\frac{1}{d_P(\overline{\Lambda}_j'(\theta_j), \overline{\Lambda}_i'(0))} + \delta \right) \log(T) \right). && (77)
\end{aligned}$$

In (73), we have performed a change-of-measure from environment ν to $\widetilde{\nu}$. In (74), we use the fact that $\{N_i(T) > (1-\gamma)T\} \supset \mathcal{A}_T$ for sufficiently large T. We have used the event $\mathcal{B}_T$ in (75), and

the relevant identity for the convex conjugates $\overline{\Lambda}_j^*$ in (76). We have used the event $\mathcal{A}_T$ in (77). We also note that $\lim_{T\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T,\mathcal{B}_T)=1$. In environment $\widetilde{\nu}$, $\lim_{T\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T)=1$ is due to the $\mathcal{M}_P$-pathwise convergence property of the algorithm π, as in (22). And $\lim_{T\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_T)=1$ is due to the same result for $\mathcal{A}_T$, together with the sample mean WLLN that comes from Assumptions 1-2 (using the upper bound part of the Gärtner-Ellis Theorem; for details, see Lemma 3.2.5 of Bucklew (2004)). From (77), taking logs and dividing by $\log(T)$, and sending $T\to\infty$ followed by $\delta\downarrow 0$, we obtain:

$$\liminf_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_i(T)>(1-\gamma)T)}{\log(T)}\geq-\sum_{j=1}^{i-1}\frac{\overline{\Lambda}_j^*(\overline{\Lambda}_j'(\theta_j))}{d_P(\overline{\Lambda}_j'(\theta_j),\overline{\Lambda}_i'(0))}.$$

This holds for any $\theta_j\in\overline{\Theta}_j$, $j\leq i-1$ such that $\overline{\Lambda}_j'(\theta_j)<\overline{\Lambda}_i'(0)$. Under Assumptions 1-2, each $\overline{\Lambda}_j'$ is an invertible mapping between $\overline{\Theta}_j$ and $\overline{\mathcal{I}}_j$ (see Theorem 26.5 of Rockafellar (1970)). Thus,

$$\liminf_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_i(T)>(1-\gamma)T)}{\log(T)}\geq-\sum_{j=1}^{i-1}\inf_{z\in\overline{\mathcal{I}}_j\,:\,z<\overline{\Lambda}_i'(0)}\frac{\overline{\Lambda}_j^*(z)}{d_P(z,\overline{\Lambda}_i'(0))}.$$

The conclusion in (23), with the infimum over $B_\gamma(T)=[\log^{1+\gamma}(T),(1-\gamma)T]$, follows from part (i) of Lemma 5 with $g(t)=\log^{1+\gamma}(t)$. □

Appendix D: Proof of Theorem 4

In the proof below, for any distributions Q and Q', we use dQ/dQ' to denote the Radon-Nikodym derivative of the absolutely continuous part of Q with respect to Q', in accordance with the Lebesgue decomposition of Q with respect to Q' (see Theorem 6.10 of Rudin (1987) for a precise statement), and we write $Q\ll Q'$ if Q is absolutely continuous with respect to Q'.

Proof of Theorem 4. Suppose there is an environment $\widetilde{\nu}=(\widetilde{P}_1,P_2,\ldots,P_K)\in\mathcal{M}^K$ for which (32) is false. Without loss of generality, let arm 2 be optimal (i.e., $\mu(P_2)=\mu_*(\widetilde{\nu})$), and suppose for sub-optimal arm 1 there exists $\epsilon\in(0,1)$ and a sequence of deterministic times $T_n\uparrow\infty$ such that for all n,

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(\frac{N_1(T_n)}{f(T_n)}\leq\frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1,\mu(P_2),\mathcal{M})}\right)\geq\epsilon. \tag{78}$$

Denote the event in (78) by $\mathcal{A}_n$. Consider any $P_1\in\mathcal{M}$ such that $\widetilde{P}_1\ll P_1$, $\mu(P_1)>\mu(P_2)$, and

$$\frac{D(\widetilde{P}_1\,\|\,P_1)}{D_{\inf}(\widetilde{P}_1,\mu(P_2),\mathcal{M})}\leq 1+\epsilon. \tag{79}$$

(Such P_1 exists or else $D_{\inf}(\widetilde{P}_1,\mu(P_2),\mathcal{M})=\infty$ and (32) would hold trivially for $\widetilde{\nu}$.) Let $\nu=(P_1,P_2,\ldots,P_K)\in\mathcal{M}^K$ so that arm 1 is now optimal, with $P_2,\ldots,P_K$ the same as in $\widetilde{\nu}$. Let $\delta>0$, and define the events:

$$\mathcal{B}_n=\left\{\left|\frac{1}{N_1(T_n)}\sum_{t=1}^{N_1(T_n)}\log\frac{dP_1}{d\widetilde{P}_1}(X_1(t))+D(\widetilde{P}_1\,\|\,P_1)\right|\leq\delta\right\}$$
$$\mathcal{C}_n=\{\exists\, i\neq 1:N_i(T_n)>T_n/K\}.$$

By a change of measure from ν to $\widetilde{\nu}$ (with an inequality due to the possibility that $P_1 \not\ll \widetilde{P}_1$),

$$\mathbb{P}_{\nu\pi}(\mathcal{C}_n) \geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}(\mathcal{C}_n)\prod_{t=1}^{N_1(T_n)}\frac{dP_1}{d\widetilde{P}_1}(X_1(t))\right] \tag{80}$$

$$\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}(\mathcal{A}_n,\mathcal{B}_n)\exp\left(\frac{1}{N_1(T_n)}\sum_{t=1}^{N_1(T_n)}\log\frac{dP_1}{d\widetilde{P}_1}(X_1(t))\cdot N_1(T_n)\right)\right] \tag{81}$$

$$\geq \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_n,\mathcal{B}_n)\cdot\exp\left(-\left(D(\widetilde{P}_1\,\|\,P_1)+\delta\right)\cdot\frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1,\mu(P_2),\mathcal{M})}f(T_n)\right), \tag{82}$$

where (81) follows from $\mathcal{C}_n \supset \mathcal{A}_n$ for large n since $f(t)=o(t)$, and (82) follows from lower bounds using $\mathcal{A}_n$ and $\mathcal{B}_n$. By Lemma 6 (see below) and the WLLN for sample means, $\lim_{n\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_n)=1$. So from (78), $\liminf_{n\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_n,\mathcal{B}_n)\geq\epsilon$. From (82), taking logs and dividing by $f(T_n)$, sending $n\to\infty$ followed by $\delta\downarrow 0$, and then applying (79), we obtain:

$$\liminf_{n\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(\mathcal{C}_n)}{f(T_n)} \geq \liminf_{n\to\infty}\frac{\log\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_n,\mathcal{B}_n)}{f(T_n)} - (1-\epsilon)\frac{D(\widetilde{P}_1\,\|\,P_1)}{D_{\inf}(\widetilde{P}_1,\mu(P_2),\mathcal{M})} \geq -(1-\epsilon^2). \tag{83}$$

Noting that $(K-1)\cdot\max_{i\neq 1}\mathbb{P}_{\nu\pi}(N_i(T_n)>T_n/K)\geq\mathbb{P}_{\nu\pi}(\mathcal{C}_n)$, we obtain:

$$\liminf_{n\to\infty}\frac{\max_{i\neq 1}\log\mathbb{P}_{\nu\pi}(N_i(T_n)>T_n/K)}{f(T_n)}\geq -(1-\epsilon^2).$$

Since $\epsilon\in(0,1)$, this violates (31) for some sub-optimal arm $i\neq 1$ (under environment ν), and thus (78) cannot be true. □

The proof of Lemma 6 is a simplification of the proof of Theorem 4.

LEMMA 6. *Under the assumptions of Theorem 4, for any environment $\nu=(P_1,\ldots,P_K)\in\mathcal{M}^K$ and each sub-optimal arm i, we have $N_i(T)\to\infty$ in $\mathbb{P}_{\nu\pi}$-probability as $T\to\infty$.*

Proof of Lemma 6. Suppose the conclusion is false for some environment $\widetilde{\nu}=(\widetilde{P}_1,P_2,\ldots,P_K)\in\mathcal{M}^K$. Without loss of generality, suppose arm 1 is sub-optimal in $\widetilde{\nu}$ and there exists $m>0$, $\epsilon>0$ and a deterministic sequence of times $T_n\uparrow\infty$ such that for all n,

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(N_1(T_n)\leq m\right)\geq\epsilon. \tag{84}$$

Denote the event in (84) by $\mathcal{A}'_n$. Consider another environment $\nu=(P_1,P_2,\ldots,P_K)\in\mathcal{M}^K$ where arm 1 is optimal (with all other arms being the same as in $\widetilde{\nu}$). Pick $L>0$ large enough so that

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(\forall\, l=1,\ldots,m:\frac{1}{l}\sum_{t=1}^{l}\log\frac{dP_1}{d\widetilde{P}_1}(X_1(t))\geq -L\right)\geq 1-\epsilon/2. \tag{85}$$

Define

$$\mathcal{B}'_n=\left\{\frac{1}{N_1(T_n)}\sum_{t=1}^{N_1(T_n)}\log\frac{dP_1}{d\widetilde{P}_1}(X_1(t))\geq -L\right\}.$$

Following the same steps from (80)-(82) but with $\mathcal{A}'_n$, $\mathcal{B}'_n$ in the place of $\mathcal{A}_n$, $\mathcal{B}_n$, respectively,

$$\mathbb{P}_{\nu\pi}(\exists\, i \neq 1 : N_i(T_n) > T_n/K) \geq \mathbb{P}_{\tilde{\nu}\pi}(\mathcal{A}'_n, \mathcal{B}'_n) \cdot \exp(-Lm)\,.$$

By (84)-(85), we have $\mathbb{P}_{\tilde{\nu}\pi}(\mathcal{A}'_n, \mathcal{B}'_n) \geq \epsilon/2$ for all n. Like previously with (83), this violates (31) for some sub-optimal arm $i \neq 1$ (under environment ν), and so (84) cannot be true. □

References

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

Ashutosh K, Nair J, Kagrecha A, Jagannathan K (2021) Bandit algorithms: letting go of logarithmic regret for statistical robustness. *International Conference on Artificial Intelligence and Statistics* .

Audibert J, Munos R, Szepesvári C (2009) Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* 410(19):1876–1902.

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

Baudry D, Gautron R, Kaufmann E, Maillard O (2021) Optimal Thompson sampling strategies for support-aware CVaR bandits. *International Conference on Machine Learning* .

Bucklew J (2004) *Introduction to Rare Event Simulation* (Springer).

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

Cappé O, Garivier A, Maillard O, Munos R, Stoltz G (2013) Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics* 41(3):1516–1541.

Cassel A, Mannor S, Zeevi A (2018) A general approach to multi-armed bandits under risk criteria. *Conference on Learning Theory* .

Cover T, Thomas J (2006) *Elements of Information Theory* (Wiley-Interscience).

Cowan W, Katehakis M (2019) Exploration–exploitation policies with almost sure, arbitrarily slow growing asymptotic regret. *Probability in the Engineering and Informational Sciences* 34(3):406–428.

Dembo A, Zeitouni O (1998) *Large Deviations Techniques and Applications* (Springer-Verlag).

Embrechts P, Klüppelberg C, Mikosch T (1997) *Modelling Extremal Events for Insurance and Finance* (Springer-Verlag).

Fan L, Glynn P (2021) Diffusion Approximations for Thompson Sampling. *arXiv:2105.09232* .

Fan L, Glynn P (2022) The Typical Behavior of Bandit Algorithms. *arXiv:2210.05660* .

Foster DJ, Gentile C, Mohri M, Zimmert J (2020) Adapting to Misspecification in Contextual Bandits. *Advances in Neural Information Processing Systems*, 11478–11489.

Galichet N, Sebag M, Teytaud O (2013) Exploration vs exploitation vs safety: risk-aware multi-armed bandits. *Asian Conference on Machine Learning* 245–260.

Garivier A, Cappé O (2011) The KL-UCB algorithm for bounded stochastic bandits and beyond. *Conference on Learning Theory* 359–376.

Ghosh A, Chowdhury SR, Gopalan A (2017) Misspecified Linear Bandits. *Association for the Advancement of Artificial Intelligence*, 3761–3767.

Kalvit A, Zeevi A (2021) A Closer Look at the Worst-case Behavior of Multi-armed Bandit Algorithms. *Advances in Neural Information Processing Systems* .

Khajonchotpanya N, Xue Y, Rujeerapaiboon N (2021) A revised approach for risk-averse multi-armed bandits under CVaR criterion. *Operations Research Letters* 49(4):465–472.

Kontoyiannis I, Meyn S (2003) Spectral theory and limit theorems for geometrically ergodic Markov processes. *The Annals of Applied Probability* 13(1):304–362.

Korda N, Kaufmann E, Munos R (2013) Thompson sampling for 1-dimensional exponential family bandits. *NeurIPS* 26.

Krishnamurthy SK, Hadad V, Athey S (2021) Adapting to misspecification in contextual bandits with offline regression oracles. *Proceedings of the 38th International Conference on Machine Learning*, 5805–5814.

Kuang X, Wager S (2021) Weak Signal Asymptotics for Sequentially Randomized Experiments. *arXiv:2101.09855* .

Lai T (1987) Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics* 15(3):1091–1114.

Lai T, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4–22.

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

Lattimore T, Szepesvari C, Weisz G (2020) Learning with Good Feature Representations in Bandits and in RL with a Generative Model. *Proceedings of the 37th International Conference on Machine Learning*, 5662–5670.

Liu Y, Devraj A, Van Roy B, Kuang X (2022) Gaussian Imagination in Bandit Learning. *arXiv:2201.01902* .

Maillard O (2013) Robust risk-averse stochastic multi-armed bandits. *International Conference on Algorithmic Learning Theory* 218–233.

Maillard O, Munos R, Stoltz G (2011) A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. *Conference on Learning Theory* 497–514.

Miller H (1961) A Convexity Property in the Theory of Random Variables Defined on a Finite Markov Chain. *The Annals of Mathematical Statistics* 32(4):1260–1270.

Moulos V, Anantharam V (2019) Optimal Chernoff and Hoeffding bounds for finite state Markov chains. *arXiv:1907.04467* .

Polanskiy Y, Wu Y (2023) *Information Theory: From Coding to Learning* (Cambridge University Press).

Prashanth L, Jagannathan K, Kolla R (2020) Concentration bounds for CVaR estimation: the cases of light-tailed and heavy-tailed distributions. *International Conference on Machine Learning* .

Rockafellar R (1970) *Convex Analysis* (Princeton University Press).

Rudin W (1987) *Real and Complex Analysis* (McGraw-Hill).

Salomon A, Audibert J (2011) Deviations of stochastic bandit regret. *International Conference on Algorithmic Learning Theory* 159–173.

Sani A, Lazaric A, Munos R (2012) Risk-aversion in multi-armed bandits. *Advances in Neural Information Processing Systems* .

Simchowitz M, Tosh C, Krishnamurthy A, Hsu DJ, Lykouris T, Dudik M, Schapire RE (2021) Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 26382–26394.

Szorenyi B, Busa-Fekete R, Weng P, Hullermeier E (2015) Qualitative multi-armed bandits: a quantile-based approach. *International Conference on Machine Learning* .

Takemura K, Ito S, Hatano D, Sumita H, Fukunaga T, Kakimura N, Kawarabayashi Ki (2021) A Parameter-Free Algorithm for Misspecified Linear Contextual Bandits . *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 3367–3375.

Tamkin A, Keramati R, Dann C, Brunskill E (2019) Distributionally-aware exploration for CVaR bandits. *Advances in Neural Information Processing Systems* .

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

Vakili S, Zhao Q (2016) Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing* 10(6):1093–1111.

Zhu Q, Tan V (2020) Thompson sampling for mean-variance bandits. *International Conference on Machine Learning* .

Zimin A, Ibsen-Jensen R, Chatterjee K (2014) Generalized risk-aversion in stochastic multi-armed bandits. *arXiv:1405.0833* .

Online Supplement for "The Fragility of Optimized Bandit Algorithms"

Appendix EC.1: Proofs for Section 3.1

For the proofs in Appendix EC.1, we will work with the natural parameterization of the exponential family in (1):

$$P_\theta(dx) = \exp\left(\theta \cdot x - \Lambda_P(\theta)\right) P(dx), \quad \theta \in \Theta_P. \tag{EC.1}$$

Then, the KL divergence between distributions P_θ and P_{θ_0} has the expression:

$$D(P_\theta \,\|\, P_{\theta_0}) = \Lambda_P(\theta_0) - \Lambda_P(\theta) - \Lambda_P'(\theta) \cdot (\theta_0 - \theta). \tag{EC.2}$$

Proof of Lemma 1. First of all, the definition of discrimination equivalence, as expressed in (6) for the exponential family with base distribution P parameterized by mean (as in (1)), is equivalent to the following statement for the same exponential family with natural parameterization (as in (EC.1)). For any $\theta_1, \theta_2 \in \Theta_P$ with $\theta_1 > \theta_2$,

$$\inf_{\theta \in \Theta_P \,:\, \theta < \theta_2} \frac{D(P_\theta \,\|\, P_{\theta_1})}{D(P_\theta \,\|\, P_{\theta_2})} = 1. \tag{EC.3}$$

Also, we recall that on Θ_P, Λ_P is strictly convex and Λ_P' is strictly increasing.

We first show the forward direction, that (EC.3) implies (9). Suppose $\inf \Theta_P > -\infty$. Note that (EC.3) implies that for any fixed $\theta_0 > \inf \Theta_P$,

$$\lim_{\theta \downarrow \inf \Theta_P} D(P_\theta \,\|\, P_{\theta_0}) = \lim_{\theta \downarrow \inf \Theta_P} \Lambda_P(\theta_0) - \Lambda_P(\theta) - \Lambda_P'(\theta) \cdot (\theta_0 - \theta) = \infty. \tag{EC.4}$$

Because $\inf \Theta_P > -\infty$ and Λ_P is strictly convex, we must have

$$\lim_{\theta \downarrow \inf \Theta_P} \Lambda_P(\theta) > -\infty, \tag{EC.5}$$

and so (EC.4) implies that

$$\lim_{\theta \downarrow \inf \Theta_P} \Lambda_P'(\theta) = -\infty. \tag{EC.6}$$

Then, taking θ_0 arbitrarily close to $\inf \Theta_P$ in (EC.4), we have for any $\epsilon > 0$:

$$\lim_{\theta \downarrow \inf \Theta_P} \Lambda_P(\theta) + \epsilon \Lambda_P'(\theta) = -\infty. \tag{EC.7}$$

So, (EC.5), (EC.6) and (EC.7) imply that

$$\lim_{\theta \downarrow \inf \Theta_P} \frac{\Lambda_P(\theta)}{\Lambda_P'(\theta)} = 0.$$

So, for any θ_1, θ_2 fixed with $\theta_1 > \theta_2 > \inf \Theta_P$, we have

$$\lim_{\theta \downarrow \inf \Theta_P} \frac{D(P_\theta \,\|\, P_{\theta_1})}{D(P_\theta \,\|\, P_{\theta_2})} = \lim_{\theta \downarrow \inf \Theta_P} \frac{\Lambda_P'(\theta) \cdot (\theta_1 - \theta)}{\Lambda_P'(\theta) \cdot (\theta_2 - \theta)} = \frac{\theta_1 - \inf \Theta_P}{\theta_2 - \inf \Theta_P} > 1,$$

which contradicts (EC.3) if $\inf \Theta_P > -\infty$. Hence, it must be that $\inf \Theta_P = -\infty$.

Given $\inf \Theta_P = -\infty$, now suppose that

$$\liminf_{\theta \to -\infty} \big(\theta \Lambda_P'(\theta) - \Lambda_P(\theta)\big) < \infty. \tag{EC.8}$$

Consider the two possible cases:

1. $\lim_{\theta \to -\infty} \Lambda_P'(\theta) = -\infty$
2. $\lim_{\theta \to -\infty} |\Lambda_P'(\theta)| < \infty$.

(Since on Θ_P, Λ_P is strictly convex and Λ_P' is strictly increasing, we cannot have $\lim_{\theta \to -\infty} \Lambda_P'(\theta) = \infty$.) In the first case, (EC.3) cannot hold because (EC.2) and (EC.8) prevent (EC.4) for $\theta_0 < 0$. In the second case, (EC.3) cannot hold because (EC.2) and (EC.8) imply that $\liminf_{\theta \to -\infty} D(P_\theta \,\|\, P_{\theta_0}) < \infty$ for any $\theta_0 \in \Theta_P$, thus preventing (EC.4). So, it must be that

$$\lim_{\theta \to -\infty} \big(\theta \Lambda_P'(\theta) - \Lambda_P(\theta)\big) = \infty.$$

Thus, the forward direction is established.

We now show the reverse direction, that (9) implies (EC.3). Starting with (9), there are again two possible cases:

1. $\lim_{\theta \to -\infty} \Lambda_P'(\theta) = -\infty$
2. $\lim_{\theta \to -\infty} |\Lambda_P'(\theta)| < \infty$.

In the first case, using the fact that $\lim_{\theta \to -\infty} D(P_\theta \,\|\, P_{\theta_0}) > 0$ for arbitrarily negative values of θ_0, together with the identity in (EC.2), we conclude that

$$\lim_{\theta \to -\infty} \frac{\Lambda_P'(\theta)}{\theta \Lambda_P'(\theta) - \Lambda_P(\theta)} = 0. \tag{EC.9}$$

Then, (EC.2), (9) and (EC.9) imply that

$$\lim_{\theta \to -\infty} \frac{D(P_\theta \,\|\, P_{\theta_1})}{D(P_\theta \,\|\, P_{\theta_2})} = \lim_{\theta \to -\infty} \frac{\theta \Lambda_P'(\theta) - \Lambda_P(\theta)}{\theta \Lambda_P'(\theta) - \Lambda_P(\theta)} = 1. \tag{EC.10}$$

In the second case, (9) directly implies (EC.9), which again leads to (EC.10). Thus, the reverse direction is established. □

Proof of Proposition 1. Since $\inf \Theta_P = -\infty$ and the support of the distributions is unbounded to the left (i.e., there is always positive probability mass to the left of any point on the real line),

as we send θ to $-\infty$, the mean $\mu(P_\theta) = \Lambda_P'(\theta)$ must also go to $-\infty$. By the definition of the convex conjugate Λ_P^*, we have for any $\theta \in \Theta_P$,

$$\Lambda_P^*(z) \geq \theta \cdot z - \Lambda_P(\theta),$$

which implies for $\theta < 0$ that

$$\lim_{z \to -\infty} \Lambda_P^*(z) = \infty.$$

Also, note that for any $\theta \in \Theta_P$,

$$\Lambda_P^*(\Lambda_P'(\theta)) = \theta \cdot \Lambda_P'(\theta) - \Lambda_P(\theta).$$

So, the desired result follows from the fact that $\lim_{\theta \to -\infty} \Lambda_P'(\theta) = -\infty$, together with Lemma 1. □

Proof of Proposition 2. Let X be a random variable with distribution P. We first address the case in which P assigns zero probability mass to the (finite) infimum of its support, which we denote by L. For $l > L$, we have by the definition of convex conjugation:

$$\begin{aligned} \Lambda_P^*(l) &= \sup_{\theta \in \Theta_P} \left(\theta \cdot l - \log \mathbb{E}[\exp(\theta X)] \right) \\ &= -\log \left(\inf_{\theta \in \Theta_P} \mathbb{E}[\exp(\theta(X - l))] \right). \end{aligned} \tag{EC.11}$$

For any $\theta \in \Theta_P$ and $l > L$,

$$0 \leq \mathbb{E}[\exp(\theta(X - l))] \leq \exp(|\theta(l - L)|) \cdot \mathbb{E}[\exp(\theta(X - L))].$$

Therefore,

$$\begin{aligned} 0 \leq \inf_{\theta \in \Theta_P} \mathbb{E}[\exp(\theta(X - l))] &\leq \exp(\left|-(l - L)^{-1}(l - L)\right|) \cdot \mathbb{E}[\exp(-(l - L)^{-1}(X - L))] \\ &= \exp(1) \cdot \mathbb{E}[\exp(-(l - L)^{-1}(X - L))], \end{aligned}$$

and since $X > L$ with probability one, we have by the Bounded Convergence Theorem, $\lim_{l \downarrow L} \mathbb{E}[\exp(-(l - L)^{-1}(X - L))] = 0$. So,

$$\lim_{l \downarrow L} \inf_{\theta \in \Theta_P} \mathbb{E}[\exp(\theta(X - l))] = 0,$$

which, by (EC.11), translates into

$$\lim_{l \downarrow L} \Lambda_P^*(l) = \infty.$$

Since

$$\lim_{\theta\to-\infty}\Lambda_P'(\theta)=L,$$

we have

$$\lim_{\theta\to-\infty}\Lambda_P^*(\Lambda_P'(\theta))=\infty, \tag{EC.12}$$

which is the equivalent representation for (9).

For the case in which P places strictly positive mass on L, fix some $\eta\in(0,1)$ such that $\mathbb{P}(X\geq L+\eta)\geq\eta$. Then, taking $m=(2/\eta)\log(1/\eta)$ and $l\in(L,L+\eta/2)$, we have

$$\begin{aligned}\inf_{\theta\geq m}\mathbb{E}[\exp(\theta(X-l))]&=\inf_{\theta\geq m}\Big\{\mathbb{E}[\exp(\theta(X-l));X\geq l]+\mathbb{E}[\exp(\theta(X-l));X<l]\Big\}\\&\geq\exp(m(L+\eta-l))\mathbb{P}(X\geq L+\eta)\\&\geq\exp\left(m\frac{\eta}{2}\right)\eta\\&=1=\mathbb{E}[\exp(0\cdot(X-l))].\end{aligned}$$

So, it suffices to take the infimum over $\theta<m$:

$$\begin{aligned}\inf_{\theta\in\Theta_P}\mathbb{E}[\exp(\theta(X-l))]&=\inf_{\theta<m}\mathbb{E}[\exp(\theta(X-l))]\\&\geq\inf_{\theta<m}\mathbb{E}[\exp(\theta(X-l));X=L]\\&=\inf_{\theta<m}\exp(\theta(L-l))\cdot\mathbb{P}(X=L)\\&=\exp(m(L-l))\cdot\mathbb{P}(X=L).\end{aligned}$$

Therefore,

$$\liminf_{l\downarrow L}\inf_{\theta\in\Theta_P}\mathbb{E}[\exp(\theta(X-l))]\geq\mathbb{P}(X=L),$$

which, by (EC.11), translates into

$$\limsup_{l\downarrow L}\Lambda_P^*(l)\leq-\log\mathbb{P}(X=L)<\infty,$$

since $\mathbb{P}(X=L)\in(0,1)$ by assumption. So, although

$$\lim_{\theta\to-\infty}\Lambda_P'(\theta)=L,$$

unlike in the case of continuous distributions, where we ended up with (EC.12), here we have

$$\limsup_{\theta\to-\infty}\Lambda_P^*(\Lambda_P'(\theta))<\infty.$$

□

Derivations for Examples 1-4.

Example 1

From (10), we have for $z_2 < z_1 \in \mathbb{R}$,

$$\inf_{z \in \mathcal{I}_P : z < z_2} \frac{d_P(z, z_1)}{d_P(z, z_2)} = \lim_{z \downarrow -\infty} \frac{(z - z_1)^2}{(z - z_2)^2} = 1.$$

So, P is discrimination equivalent.

Example 2

The CGF is $\theta \mapsto \log((e^\theta - 1)/\theta)$. Let $\theta_1 = \theta_P(z_1)$, $\theta_2 = \theta_P(z_2)$ be the tilting parameters corresponding to the means z_1, z_2, with $0 < z_2 < z_1 < 1$. Note that the mean $z \downarrow 0$ corresponds to $\theta_P(z) \downarrow -\infty$. From (10) and using (3), we have

$$\begin{aligned}\inf_{z \in \mathcal{I}_P : z < z_2} \frac{d_P(z, z_1)}{d_P(z, z_2)} &= \lim_{\theta \to -\infty} \frac{\log\left(\frac{e^{\theta_1} - 1}{\theta_1}\right) - \log\left(\frac{e^\theta - 1}{\theta}\right) - \frac{e^\theta(\theta - 1) + 1}{(e^\theta - 1)\theta}(\theta_1 - \theta)}{\log\left(\frac{e^{\theta_2} - 1}{\theta_2}\right) - \log\left(\frac{e^\theta - 1}{\theta}\right) - \frac{e^\theta(\theta - 1) + 1}{(e^\theta - 1)\theta}(\theta_2 - \theta)} \\ &= 1.\end{aligned}$$

So, P is discrimination equivalent.

Example 3

From (10), we have for $0 < z_2 < z_1 < 1$,

$$\begin{aligned}\inf_{z \in \mathcal{I}_P : z < z_2} \frac{d_P(z, z_1)}{d_P(z, z_2)} &= \lim_{z \downarrow 0} \frac{z \log\left(\frac{z}{z_1}\right) + (1 - z) \log\left(\frac{1 - z}{1 - z_1}\right)}{z \log\left(\frac{z}{z_2}\right) + (1 - z) \log\left(\frac{1 - z}{1 - z_2}\right)} \\ &= \frac{\log(1 - z_1)}{\log(1 - z_2)}.\end{aligned}$$

So, P is not discrimination equivalent.

Example 4

The density of P^z is $x \mapsto (\theta_P(z) + 1) e^{(\theta_P(z)+1)x} \mathbb{I}(x \le 0)$, where $\theta_P(z) \in (-1, \infty)$ is the tilting parameter corresponding to mean $z \in \mathbb{R}$. The KL-divergence has the form:

$$\begin{aligned}d_P(z, z') &= \int_{-\infty}^{0} (\theta_P(z) + 1) e^{(\theta_P(z)+1)x} \left(\log\left(\frac{\theta_P(z) + 1}{\theta_P(z') + 1}\right) + (\theta_P(z) - \theta_P(z'))x \right) dx \\ &= -\frac{\theta_P(z) - \theta_P(z')}{\theta_P(z) + 1} + \log\left(\frac{\theta_P(z) + 1}{\theta_P(z') + 1}\right).\end{aligned}$$

Let $\theta_1 = \theta_P(z_1)$, $\theta_2 = \theta_P(z_2)$ be the tilting parameters corresponding to the means z_1, z_2, with $z_2 < z_1 < 0$. Note that the mean $z \downarrow -\infty$ corresponds to $\theta_P(z) \downarrow -1$. From (10), we have

$$\inf_{z \in \mathcal{I}_P : z < z_2} \frac{d_P(z, z_1)}{d_P(z, z_2)} = \lim_{\theta \downarrow -1} \frac{-\frac{\theta - \theta_1}{\theta + 1} + \log\left(\frac{\theta + 1}{\theta_1 + 1}\right)}{-\frac{\theta - \theta_2}{\theta + 1} + \log\left(\frac{\theta + 1}{\theta_2 + 1}\right)}$$

$$
\begin{aligned}
&= \frac{\theta_1 - \theta + (\theta+1)\left(\log(\theta+1) - \log(\theta_1+1)\right)}{\theta_2 - \theta + (\theta+1)\left(\log(\theta+1) - \log(\theta_2+1)\right)} \\
&= \frac{\theta_1+1}{\theta_2+1} \\
&= \frac{z_2}{z_1}.
\end{aligned}
$$

So, P is not discrimination equivalent.

Appendix EC.2: Proofs for Sections 3.2-3.3

Proof of Proposition 3. Let i be any sub-optimal arm. From the lower bounds in (47) in the proof of Theorem 1, there exists $a>0$ such that for all $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ and T sufficiently large,

$$
\begin{aligned}
T^{-a} &\le \mathbb{P}_{\nu\pi}(N_i(T) > (1-\gamma)T) \\
&\le \mathbb{P}_{\nu\pi}(N_i(T) > x).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
0 &\le \mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_i(T) - \mu_i| > \epsilon \mid N_i(T) > x\right) \\
&\le \frac{\mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_i(T) - \mu_i| > \epsilon,\ N_i(T) > \log^{1+\gamma}(T)\right)}{\mathbb{P}_{\nu\pi}\left(N_i(T) > (1-\gamma)T\right)} \\
&\le \frac{1}{\mathbb{P}_{\nu\pi}\left(N_i(T) > (1-\gamma)T\right)} \sum_{k=\lceil \log^{1+\gamma}(T)\rceil}^{\infty} \mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_i(T) - \mu_i| > \epsilon,\ N_i(T) = k\right) \\
&\le T^a \left(\frac{\exp\left(-\log^{1+\gamma}(T) \cdot d_P(\mu_i+\epsilon, \mu_i)\right)}{1 - \exp\left(-d_P(\mu_i+\epsilon, \mu_i)\right)} + \frac{\exp\left(-\log^{1+\gamma}(T) \cdot d_P(\mu_i-\epsilon, \mu_i)\right)}{1 - \exp\left(-d_P(\mu_i-\epsilon, \mu_i)\right)} \right),
\end{aligned}
$$

where the last inequality follows from a Chernoff bound. So, $\mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_i(T) - \mu_i| > \epsilon \mid N_i(T) > x\right) \to 0$ uniformly for $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ as $T \to \infty$, which yields the desired result. □

Verification of (67) in Proof of Theorem 2. With the natural parameterization of an exponential family P_θ, $\theta \in \Theta_P$, as in (EC.1), with KL divergence as in (EC.2), we have:

$$
\frac{d}{d\theta} D(P_\theta \,\|\, P_{\theta_0}) = -\Lambda_P''(\theta)(\theta_0 - \theta).
$$

Denote $\theta_1 := \theta_P(\mu_1)$ and $\theta_2 := \theta_P(\mu_2)$ (with $\theta_P(\cdot)$ as defined in the parameterization by mean in (1)), so that $\theta_2 < \theta_1$. Let $\epsilon > 0$ such that $\theta_2 + \epsilon < \theta_1$. Then,

$$
\frac{d}{d\theta} \frac{D(P_\theta \,\|\, P_{\theta_1})}{D(P_\theta \,\|\, P_{\theta_2+\epsilon})} = \frac{\Lambda_P''(\theta)}{D(P_\theta \,\|\, P_{\theta_2+\epsilon})^2} \left(\underbrace{D(P_\theta \,\|\, P_{\theta_1})(\theta_2+\epsilon-\theta) - D(P_\theta \,\|\, P_{\theta_2+\epsilon})(\theta_1-\theta)}_{:=\xi(\theta)} \right).
$$

Note that $\xi(\theta_2+\epsilon) = 0$ and $\xi'(\theta) = D(P_\theta \,\|\, P_{\theta_2+\epsilon}) - D(P_\theta \,\|\, P_{\theta_1})$ for $\theta < \theta_2 + \epsilon$. So, $\xi'(\theta) < 0$, and thus $\xi(\theta) > 0$ for $\theta < \theta_2 + \epsilon$. From this, together with the fact that $\Lambda_P''(\theta) \ge 0$ for all θ, we conclude that $\theta \mapsto D(P_\theta \,\|\, P_{\theta_1}) / D(P_\theta \,\|\, P_{\theta_2+\epsilon})$ is monotone increasing for $\theta < \theta_2 + \epsilon$.

Let $\delta > 0$ such that $\mu_2 + \delta < \mu_1$. Since $z \mapsto \theta_P(z)$ is monotone increasing, $z \mapsto d_P(z,\mu_1)/d_P(z,\mu_2+\delta)$ must also be monotone increasing for $z < \mu_2 + \delta$. So for any $\delta > 0$,

$$\inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} = \inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)}.$$

Since for $z < \mu_2$, $\delta \mapsto d_P(z,\mu_1)/d_P(z,\mu_2+\delta)$ is monotone decreasing, it must also be that $\delta \mapsto \inf_{z<\mu_2} d_P(z,\mu_1)/d_P(z,\mu_2+\delta)$ is monotone decreasing. Therefore,

$$\lim_{\delta \downarrow 0} \inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} = \sup_{\delta>0} \inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)}.$$

Finally, since both $z \mapsto d_P(z,\mu_1)/d_P(z,\mu_2+\delta)$ and $\delta \mapsto d_P(z,\mu_1)/d_P(z,\mu_2+\delta)$ are monotone, and thus are both quasi-convex and quasi-concave, Sion's Minimax Theorem yields:

$$\sup_{\delta>0} \inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} = \inf_{z<\mu_2} \sup_{\delta>0} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} = \inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2)}.$$

□

Appendix EC.3: Proofs for Section 4.2

Proof of Proposition 4. The proof follows the approach we have taken to establish previous results. We use Lemma 5 with $g(t) = \log^{1+\gamma}(t)$. In the context of part (i) of Lemma 5, the proof of the lower bound part for (19) follows from Theorem 3 (which uses Lemma 3). In the context of part (ii) of Lemma 5, the upper bound part for (19) can be established using the same arguments in the proof of Theorem 2; see Appendix B. The uniform convergence result then follows from part (iii) of Lemma 5.

Without loss of generality, suppose that $\mu(Q_1) > \mu(Q_2) > \cdots > \mu(Q_K)$ (i.e., $r(i) = i$ for all $i \in [K]$). Here, the only thing that needs to be checked is the analog of (67):

$$\lim_{\delta \downarrow 0} \inf_{z<\mu(Q_2)+\delta} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2)+\delta)} = \inf_{z<\mu(Q_2)} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2))}. \tag{EC.13}$$

(Below, we check (EC.13) for Q_1 and Q_2. The same arguments apply for the other combinations of Q_i, $i \geq 3$ and Q_j, $j \leq i-1$.) First, there exists a fixed $\eta > 0$ (depending on Q_1 and Q_2) such that for all $\delta > 0$ sufficiently small, we have both:

$$\inf_{z<\mu(Q_2)+\delta} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2)+\delta)} = \inf_{z<\mu(Q_2)-\eta} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2)+\delta)}, \tag{EC.14}$$

$$\inf_{z<\mu(Q_2)} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2))} = \inf_{z<\mu(Q_2)-\eta} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2))}. \tag{EC.15}$$

Note that

$$z \mapsto \frac{d_P(z,\mu(Q_2))}{d_P(z,\mu(Q_2)+\delta)}$$

is monotone decreasing for $z < \mu(Q_2)$, which we deduce from the verification of (67) in proof of Theorem 2 in Appendix EC.2. Also, we have:

$$\lim_{\delta \downarrow 0} \sup_{z<\mu(Q_2)-\eta} \frac{d_P(z,\mu(Q_2))}{d_P(z,\mu(Q_2)+\delta)} = \sup_{\delta>0} \sup_{z<\mu(Q_2)-\eta} \frac{d_P(z,\mu(Q_2))}{d_P(z,\mu(Q_2)+\delta)} = 1, \tag{EC.16}$$

$$\lim_{\delta \downarrow 0} \frac{d_P(\mu(Q_2)-\eta,\mu(Q_2))}{d_P(\mu(Q_2)-\eta,\mu(Q_2)+\delta)} = 1. \tag{EC.17}$$

The monotonicity property, together with (EC.16)-(EC.17), imply uniform convergence for $z < \mu(P_2) - \eta$:

$$\lim_{\delta \downarrow 0} \sup_{z<\mu(Q_2)-\eta} \left| \frac{d_P(z,\mu(Q_2))}{d_P(z,\mu(Q_2)+\delta)} - 1 \right| = 0. \tag{EC.18}$$

For any $\epsilon \in (0,1)$, using (EC.18), we have for sufficiently small $\delta > 0$:

$$(1-\epsilon) \inf_{z<\mu(Q_2)-\eta} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2))} \le \inf_{z<\mu(Q_2)-\eta} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2)+\delta)} \cdot \frac{d_P(z,\mu(Q_2))}{d_P(z,\mu(Q_2))} \tag{EC.19}$$

$$\le (1+\epsilon) \inf_{z<\mu(Q_2)-\eta} \frac{d_{Q_1}(z,\mu(Q_1))}{d_P(z,\mu(Q_2))}. \tag{EC.20}$$

In (EC.19)-(EC.20), sending $\delta \downarrow 0$, followed by $\epsilon \downarrow 0$, and then using (EC.14)-(EC.15), we obtain (EC.13). □

Proof of Corollary 2. Let ν consist of two Gaussian reward distributions with variance σ_0^2, and μ_1 and μ_2 as the means for arms 1 and 2, respectively. Without loss of generality, suppose that $\mu_1 > \mu_2$ (i.e., $r(i) = i$ for $i = 1, 2$). We again use Lemma 5 with $g(t) = \log^{1+\gamma}(t)$. In the context of part (i) of Lemma 5, the proof of the lower bound part for (19) follows from Theorem 3 (which uses Lemma 3). In the context of part (ii) of Lemma 5, the upper bound part:

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > \log^{1+\gamma}(T))}{\log(\log^{1+\gamma}(T))} \le -\frac{\sigma^2}{\sigma_0^2}, \tag{EC.21}$$

actually follows from the proof of the upper bound part of Theorem 2. In the Gaussian setting, the proof is substantially simpler, and so for future reference, we provide it below. The uniform convergence result then follows from part (iii) of Lemma 5. □

Verification of (EC.21) in Proof of Corollary 2. Let $x_T = \lfloor \log^{1+\gamma}(T) \rfloor$ with fixed $\gamma \in (0,1)$. Let $\Delta = \mu_1 - \mu_2 > 0$. As in the proof of Theorem 2, we have:

$$\begin{aligned}
&\mathbb{P}_{\nu\pi}\left(N_2(T) > x_T\right) \\
&\le \mathbb{P}_{\nu\pi}\left(\exists\, t \in (\tau_2(x_T), T] \ \text{ s.t. } \ \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(t-1)}{N_1(t-1)}} \le \widehat{\mu}_2(t-1) + \sqrt{\frac{2\sigma^2 \log(t-1)}{N_2(t-1)}} \right)
\end{aligned}$$

$$\leq \mathbb{P}_{\nu\pi}\left(\exists\, t\in(x_T,T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1)+\sqrt{\frac{2\sigma^2\log(x_T)}{N_1(t-1)}}\leq \widehat{\mu}_2(\tau_2(x_T))+\sqrt{\frac{2\sigma^2\log(T)}{x_T}}\right)$$

$$\leq \mathbb{P}_{\nu\pi}\left(\exists\, t\in(x_T,T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1)+\sqrt{\frac{2\sigma^2\log(x_T)}{N_1(t-1)}}\leq \mu_2+\frac{\Delta}{2}\right) \tag{EC.22}$$

$$+\,\mathbb{P}_{\nu\pi}\left(\widehat{\mu}_2(\tau_2(x_T))+\sqrt{\frac{2\sigma^2\log(T)}{x_T}}>\mu_2+\frac{\Delta}{2}\right). \tag{EC.23}$$

For the term in (EC.22), we have

$$(EC.22)=\mathbb{P}_{\nu\pi}\left(\exists\, t\in(x_T,T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1)+\sqrt{\frac{2\sigma^2\log(x_T)}{N_1(t-1)}}\leq \mu_1-\frac{\Delta}{2}\right)$$

$$\leq \sum_{m=1}^{\infty}\mathbb{P}_{\nu\pi}\left(\frac{1}{m}\sum_{l=1}^{m}X_1(l)\leq \mu_1-\sqrt{\frac{2\sigma^2\log(x_T)}{m}}-\frac{\Delta}{2}\right) \tag{EC.24}$$

$$\leq \sum_{m=1}^{\infty}\exp\left(-\frac{m}{2\sigma_0^2}\left(\sqrt{\frac{2\sigma^2\log(x_T)}{m}}+\frac{\Delta}{2}\right)^2\right) \tag{EC.25}$$

$$= x_T^{-\sigma^2/\sigma_0^2}\cdot\sum_{m=1}^{\infty}\exp\left(-\frac{\sqrt{m\sigma^2\log(x_T)}\Delta}{\sqrt{2}\sigma_0^2}-\frac{m\Delta^2}{8\sigma_0^2}\right)$$

$$\leq x_T^{-\sigma^2/\sigma_0^2}\cdot\sum_{m=1}^{\infty}\exp\left(-\frac{\sqrt{m}\sigma\Delta}{\sqrt{2}\sigma_0^2}-\frac{m\Delta^2}{8\sigma_0^2}\right) \qquad (\text{for } T\geq 16), \tag{EC.26}$$

where to obtain (EC.24), we have used a union bound over all possible values of $N_1(t)$, $t\geq 1$, and (EC.25) follows from a Chernoff bound.

For the term in (EC.23), we have for sufficiently large T,

$$\sqrt{\frac{2\sigma^2\log(T)}{x_T}}<\frac{\Delta}{4}.$$

So, for sufficiently large T,

$$(EC.23)\leq \mathbb{P}_{\nu\pi}\left(\frac{1}{x_T}\sum_{t=1}^{x_T}X_2(t)>\mu_2+\frac{\Delta}{4}\right)$$

$$\leq \exp\left(-x_T\cdot\frac{\Delta^2}{32\sigma_0^2}\right), \tag{EC.27}$$

where (EC.27) follows from a Chernoff bound.

Putting together (EC.22), (EC.26) and (EC.23), (EC.27), we have established the desired result:

$$\limsup_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}\left(N_2(T)>x_T\right)}{\log(x_T)}\leq -\frac{\sigma^2}{\sigma_0^2}.$$

□

Appendix EC.4: Proofs for Section 4.3

Proof of Lemma 3. This proof is an extension and simplification of Propositions 7-8 of Cowan and Katehakis (2019).

We restrict our attention to sample paths ω belonging to

$$\left\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} X_i(t) = \mu_i,\ i \in [K]\right\}. \tag{EC.28}$$

Without loss of generality, suppose that arm 1 is the unique optimal arm, i.e., $\mu_1 > \max_{i\geq 2}\mu_i$. The UCB index for arm i at time $t+1$ is:

$$U_i(t) = \sup\left\{z \in \mathcal{I}_P : d_P(\widehat{\mu}_i(t), z) \leq \frac{f(t)}{N_i(t)}\right\}, \tag{EC.29}$$

where, as defined previously, $\widehat{\mu}_i(t) = \frac{1}{N_i(t)}\sum_{s=1}^{N_i(t)} X_i(s)$. As discussed in the proof of Theorem 2, $f(t)$ is a design choice. For KL-UCB, choices include $f(t) = \log(t)$, $f(t) = \log(t) + 3\log\log(t)$ and $f(t) = \log(1 + t\log^2(t))$. We will leave the particular form for $f(t)$ unspecified in developing this proof. This proof holds for any regularly varying and increasing function $f : (1,\infty) \to (0,\infty)$ satisfying $\lim_{t\to\infty} f(t) = \infty$ and $f(t) = o(t)$.

We begin with the upper bound part of the proof. Consider sub-optimal arm $i \geq 2$, and let $\delta \in (0, (\mu_1 - \mu_i)/2)$. We have

$$N_i(T) = 1 + \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ U_i(t) \geq \mu_1 - \delta,\ \widehat{\mu}_i(t) \leq \mu_i + \delta\right) \tag{EC.30}$$

$$+ \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ U_i(t) \geq \mu_1 - \delta,\ \widehat{\mu}_i(t) > \mu_i + \delta\right) \tag{EC.31}$$

$$+ \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ U_i(t) < \mu_1 - \delta\right), \tag{EC.32}$$

where $A(t)$ is the arm played by the algorithm at time t.

The first sum is upper bounded via:

$$(EC.30) \leq \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ d_P(\mu_i + \delta, \mu_1 - \delta) \leq \frac{f(t)}{N_i(t)}\right) \tag{EC.33}$$

$$\leq \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ N_i(t) \leq \frac{f(T)}{d_P(\mu_i + \delta, \mu_1 - \delta)}\right)$$

$$\leq \frac{f(T)}{d_P(\mu_i + \delta, \mu_1 - \delta)} + 1. \tag{EC.34}$$

The bound in (EC.33) holds due to the events $U_i(t) \geq \mu_1 - \delta$ and $\widehat{\mu}_i(t) \leq \mu_i + \delta$ and the definition of the index in (EC.29).

The second sum is upper bounded via:

$$(EC.31) \leq \sum_{t=K}^{\infty} \mathbb{I}\left(A(t+1)=i,\, \widehat{\mu}_i(t) > \mu_i + \delta\right). \tag{EC.35}$$

On sample paths in (EC.28), the indicators on the right side of (EC.35) can equal 1 for only finitely many t. (For each 1 in the sum, arm i is played an additional time, and an additional sample is incorporated into $\widehat{\mu}_i(t)$.)

The third sum is upper bounded via:

$$\begin{aligned}(EC.32) &\leq \sum_{t=K}^{\infty} \mathbb{I}\left(A(t+1)=i,\, U_1(t) \leq U_i(t) < \mu_1 - \delta\right) \\ &\leq \sum_{t=K}^{\infty} \mathbb{I}\left(U_1(t) < \mu_1 - \delta\right). \end{aligned} \tag{EC.36}$$

On sample paths in (EC.28), the indicators on the right side of (EC.36) can equal 1 only for finitely many t. (As $t \to \infty$, either $N_1(t)$ increases to infinity or remains bounded uniformly in t. In the first case, $\widehat{\mu}_1(t) \to \mu_1$, and so for t sufficiently large, $U_1(t) \geq \widehat{\mu}_1(t) > \mu_1 - \delta/2$. In the second case, $f(t)$ in (EC.29) increases without bound, and so $U_1(t)$ also increases without bound, with $U_1(t) > \mu_1$ for all t sufficiently large.)

Putting together (EC.34)-(EC.36), and sending $T \to \infty$ followed by $\delta \downarrow 0$, we have for each sub-optimal arm $i \geq 2$,

$$\limsup_{T\to\infty} \frac{N_i(T)}{f(T)} \leq \frac{1}{d_P(\mu_i, \mu_1)}. \tag{EC.37}$$

Therefore, for the optimal arm 1,

$$\lim_{T\to\infty} \frac{N_1(T)}{T} = 1, \tag{EC.38}$$

which, by the form of the index in (EC.29), then implies:

$$\lim_{t\to\infty} U_1(t) = \mu_1. \tag{EC.39}$$

Then, (EC.38) and (EC.39) imply that for each sub-optimal arm $i \geq 2$,

$$\lim_{T\to\infty} N_i(T) = \infty. \tag{EC.40}$$

(If (EC.40) is not true for some sub-optimal arm j, then since the term $f(t)$ grows without bound in the index (EC.29), we would eventually have $U_j(t) > \mu_1 + \epsilon > U_1(t)$ for some $\epsilon > 0$ and all t sufficiently large. This would prevent arm 1 from being played for all t sufficiently large, thereby contradicting (EC.38).)

We now develop the lower bound parts of the proof. As defined previously, for any positive integer m, we use $\tau_1(m)$ to denote the time of the m-th play of arm 1. So, for each sub-optimal arm $i \geq 2$,

$$U_1(\tau_1(m)-1) > U_i(\tau_1(m)-1). \tag{EC.41}$$

Let $\delta > 0$. We have for m sufficiently large,

$$\begin{aligned}
\max_{t\in[\tau_1(m),\tau_1(m+1)]} \frac{f(t)}{N_i(t)} &\leq \frac{f(\tau_1(m+1))}{N_i(\tau_1(m)-1)} \\
&= \frac{f(\tau_1(m+1))}{f(\tau_1(m)-1)} \frac{f(\tau_1(m)-1)}{N_i(\tau_1(m)-1)} \\
&\leq (1+\delta)\frac{f(\tau_1(m)-1)}{N_i(\tau_1(m)-1)} && \text{(EC.42)} \\
&\leq (1+\delta) d_P(\mu_i-\delta, U_i(\tau_1(m)-1)) && \text{(EC.43)} \\
&\leq (1+\delta) d_P(\mu_i-\delta, U_1(\tau_1(m)-1)) && \text{(EC.44)} \\
&\leq (1+\delta) d_P(\mu_i-\delta, \mu_1+\delta). && \text{(EC.45)}
\end{aligned}$$

Note that (EC.42) is due to (EC.38), (EC.43) is due to $\lim_{t\to\infty} \widehat{\mu}_i(t) = \mu_i$ for each sub-optimal arm $i \geq 2$ and the form of the index in (EC.29), (EC.44) is due to (EC.41), and (EC.45) is due to (EC.39). From (EC.45), we obtain

$$\liminf_{T\to\infty} \frac{N_i(T)}{f(T)} \geq \frac{1}{d_P(\mu_i,\mu_1)},$$

which together with (EC.37), completes the proof. □

Appendix EC.5: Proofs for Section 6.1

Proof of (34) in Proposition 7. Without loss of generality, suppose that $\mu_1 > \mu_2 > \cdots > \mu_K$ (i.e., $r(i) = i$ for all $1 \leq i \leq K$) for the environment ν. To simplify notation, for the right side of (34), we use the shorthand:

$$c_i(\nu) := \sum_{j=1}^{i-1} \inf_{z\in\mathcal{I}_P : z<\mu_i} \frac{d_P(z,\mu_j)}{d_P(z,\mu_i)}.$$

Recall that we consider any (fixed) $\gamma \in (0, (1/\lambda)-1)$. Let $B_\gamma(T) = [f^{1+\gamma}(T), (1-\gamma)T]$. To establish (34), we will show for any sub-optimal arm i:

$$\liminf_{T\to\infty} \inf_{x\in B_\gamma(T)} \frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > x)}{f(x)} \geq -c_i(\nu), \tag{EC.46}$$

$$\limsup_{T\to\infty} \sup_{x\in B_\gamma(T)} \frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > x)}{f(x)} \leq -c_i(\nu). \tag{EC.47}$$

This is analogous to the approach taken in Lemma 5, where parts (i) and (ii) were used to establish part (iii). However, here we must handle any regularly varying function f satisfying $\liminf_{t\to\infty} f(t)/\log(t) \geq 1$ and $f(t) = o(t^\lambda)$ for some $\lambda \in (0,1)$ (instead of just $x \mapsto \log(x)$, as in Lemma 5).

Proof of (EC.46)

Let $\gamma' \in (0,\gamma)$. We follow the proof of Theorem 1 (in Appendix A) with three changes. First, we use γ' instead of γ. Second, we replace $\log(T)$ by $f(T)$ everywhere. Third, in place of the WLLN provided by Lemma 2, we use the WLLN derived from the SLLN of Lemma 3 (see Appendix EC.4), which continues to hold for regularly varying and strictly increasing functions f satisfying $\liminf_{t\to\infty} f(t)/\log(t) \geq 1$ and $f(t) = o(t^\lambda)$ for some $\lambda \in (0,1)$. Running through the proof of Theorem 1 with these three changes, instead of obtaining (47), we obtain:

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > (1-\gamma')T)}{f(T)} \geq -c_i(\nu).$$

Since f is regularly varying, there exists $a \in (0,1)$ such that

$$\lim_{T\to\infty} \frac{f((1-\gamma')T)}{f(T)} = (1-\gamma')^a.$$

Therefore,

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > (1-\gamma')T)}{f((1-\gamma')T)} \geq -c_i(\nu)(1-\gamma')^{-a}. \tag{EC.48}$$

Since $t \mapsto N_i(t)$ is non-decreasing with $\mathbb{P}_{\nu\pi}$-probability one, we have for sufficiently large T and all $x \in [f^{1+\gamma}(T), (1-\gamma')T]$,

$$\mathbb{P}_{\nu\pi}(N_i(T) > x) \geq \mathbb{P}_{\nu\pi}(N_i(\lceil x/(1-\gamma')\rceil) > x).$$

So, for any $\epsilon > 0$, we have for sufficiently large T,

$$\frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > x)}{f(x)} \geq \frac{\log \mathbb{P}_{\nu\pi}(N_i(\lceil x/(1-\gamma')\rceil) > x)}{f(x)} \tag{EC.49}$$

$$\geq -c_i(\nu)(1-\gamma')^{-a}(1+\epsilon), \tag{EC.50}$$

uniformly for all $x \in B_\gamma(T) = [f^{1+\gamma}(T), (1-\gamma)T] \subset [f^{1+\gamma}(T), (1-\gamma')T]$, where (EC.50) follows from the convergence result in (EC.48). Then, (EC.46) is established by taking the infimum over $x \in B_\gamma(T)$ on the left side of (EC.49), sending $T \to \infty$, and then $\epsilon \downarrow 0$ and $\gamma' \downarrow 0$.

Proof of (EC.47)

Let $g(t) = f^{1+\gamma}(t)$. Using the upper bound part of the proof of Theorem 2 (in Appendix B), we obtain:

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}\left(N_i(T) > g(T)\right)}{f(g(T))} \leq -c_i(\nu). \tag{EC.51}$$

The function g is strictly increasing, so it has an inverse g^{-1} (defined on the range of g), which is also strictly increasing. Since $t \mapsto N_i(t)$ is non-decreasing with $\mathbb{P}_{\nu\pi}$-probability one, we have for sufficiently large T and all $x \in B_\gamma(T) = [g(T), (1-\gamma)T]$,

$$\mathbb{P}_{\nu\pi}(N_i(\lfloor g^{-1}(x) \rfloor) > x) \geq \mathbb{P}_{\nu\pi}(N_i(T) > x).$$

Thus, for any $\epsilon \in (0,1)$, we have for sufficiently large T,

$$-c_i(\nu)(1-\epsilon) \geq \frac{\log \mathbb{P}_{\nu\pi}(N_i(\lfloor g^{-1}(x) \rfloor) > x)}{f(x)} \tag{EC.52}$$

$$\geq \frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > x)}{f(x)}, \tag{EC.53}$$

uniformly for all $x \in B_\gamma(T)$, where (EC.52) follows from the convergence result in (EC.51). Then, (EC.47) is established by taking the supremum over $x \in B_\gamma(T)$ on the right side of (EC.53), sending $T \to \infty$, and then $\epsilon \downarrow 0$. $\square$

Proof of (35) in Proposition 7. We adapt the proof of Theorem 10.6 (page 116) of Lattimore and Szepesvári (2020), with only a few slight modifications. Recall from the proof of Theorem 2 (in Appendix B) that $\tau_i(n)$ denotes the time of the n-th play of arm i. Without loss of generality, suppose arm 1 is optimal for the environment ν under consideration. Let $i \geq 2$, and let $\epsilon_1, \epsilon_2 \in (0, (\mu_1 - \mu_i)/2)$. Define:

$$\xi_1 = \min\left\{ t : \max_{1 \leq s \leq T} \left(\underline{d}_P(\widehat{\mu}_1(\tau_1(s)), \mu_1 - \epsilon_2) - \frac{f(t)}{s} \right) \leq 0 \right\} \tag{EC.54}$$

$$\kappa_i = \sum_{s=1}^{T} \mathbb{I}\left(d_P(\widehat{\mu}_i(\tau_i(s)), \mu_1 - \epsilon_2) \leq \frac{f(T)}{s} \right), \tag{EC.55}$$

where we use the notation $\underline{d}_P(x,y) := d_P(x,y)\mathbb{I}(x \leq y)$. Then,

$$\begin{aligned}
\mathbb{E}_{\nu\pi}[N_i(T)] &= \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T} \mathbb{I}(A(t) = i) \right] \\
&\leq \mathbb{E}_{\nu\pi}[\xi_1] + \mathbb{E}_{\nu\pi}\left[\sum_{t=\xi_1+1}^{T} \mathbb{I}(A(t) = i) \right] \\
&\leq \mathbb{E}_{\nu\pi}[\xi_1] + \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{T} \mathbb{I}\left(A(t) = i,\, d_P(\widehat{\mu}_i(t-1), \mu_1 - \epsilon_2) \leq \frac{f(t)}{N_i(t-1)} \right) \right] && \text{(EC.56)} \\
&\leq \mathbb{E}_{\nu\pi}[\xi_1] + \mathbb{E}_{\nu\pi}[\kappa_i] && \text{(EC.57)} \\
&\leq C(\mu_1, \epsilon_2) + \frac{f(T)}{d_P(\mu_i + \epsilon_1, \mu_1 - \epsilon_2)} + C'(\mu_i, \epsilon_1). && \text{(EC.58)}
\end{aligned}$$

Note that (EC.56) follows from the observation that for time periods after ξ_1, the UCB index for arm 1 must be at least as large as $\mu_1 - \epsilon_2$, and so if arm i is played, then the UCB index for arm i

must also be at least as large as $\mu_1 - \epsilon_2$. Also, (EC.57) follows directly from the definition of κ_i in (EC.55). And (EC.58) follows from Lemmas EC.1 and EC.2. Therefore,

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{f(T)} \leq \frac{1}{d_P(\mu_i,\mu_1)}.$$

The matching asymptotic lower bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{f(T)} \geq \frac{1}{d_P(\mu_i,\mu_1)}$$

is directly obtained from the WLLN derived from the SLLN of Lemma 3 (see Appendix EC.4), together with Markov's inequality. □

Lemmas EC.1 and EC.2 are directly from Chapter 10 of Lattimore and Szepesvári (2020), and are included here for convenience in referencing. Lemma EC.3 is directly from the Appendix of Cappé et al. (2013), and is included here also for convenience in referencing.

LEMMA EC.1 **(Lemma 10.7 from Lattimore and Szepesvári (2020))**. *Let* $X_1,\ldots,X_T$ *be independent random variables from* P^μ*, and let* $\overline{\mu}_s = s^{-1}\sum_{j=1}^s X_j$*. Let* $\epsilon > 0$ *and define*

$$\xi = \min\left\{t : \max_{1\leq s\leq T}\left(\underline{d}_P(\overline{\mu}_s, \mu-\epsilon) - \frac{f(t)}{s}\right) \leq 0\right\}, \tag{EC.59}$$

where $f(t) \geq \log(1+t\log^2(t))$ *for sufficiently large* t*. Then,* $\mathbb{E}[\xi] \leq C(\mu,\epsilon)$*, for some finite constant* $C(\mu,\epsilon)$ *depending on* μ *and* ϵ*.*

Proof of Lemma EC.1. Define $\sigma_P^2(\mu_1,\mu_2) := \max_{z\in[l,u]} \mathrm{Var}_{X\sim P^z}(X)$, with $l = \min(\mu_1,\mu_2)$ and $u = \max(\mu_1,\mu_2)$. Then, we have

$$\begin{aligned}
\mathbb{P}(\xi > t) &\leq \mathbb{P}\left(\exists 1 \leq s \leq T : \underline{d}_P(\overline{\mu}_s, \mu-\epsilon) > \frac{f(t)}{s}\right) \\
&\leq \sum_{s=1}^T \mathbb{P}\left(\underline{d}_P(\overline{\mu}_s, \mu-\epsilon) > \frac{f(t)}{s}\right) \\
&= \sum_{s=1}^T \mathbb{P}\left(d_P(\overline{\mu}_s, \mu-\epsilon) > \frac{f(t)}{s}, \overline{\mu}_s < \mu-\epsilon\right) \\
&\leq \sum_{s=1}^T \mathbb{P}\left(d_P(\overline{\mu}_s, \mu) - d_P(\mu-\epsilon,\mu) > \frac{f(t)}{s}, \overline{\mu}_s < \mu-\epsilon\right) && \text{(EC.60)} \\
&\leq \sum_{s=1}^T \mathbb{P}\left(d_P(\overline{\mu}_s, \mu) > \frac{f(t)}{s} + \frac{\epsilon^2}{2\sigma_P^2(\mu-\epsilon,\mu)}, \overline{\mu}_s < \mu\right) && \text{(EC.61)} \\
&\leq \sum_{s=1}^T \exp\left(-s\left(\frac{f(t)}{s} + \frac{\epsilon^2}{2\sigma_P^2(\mu-\epsilon,\mu)}\right)\right) && \text{(EC.62)} \\
&\leq (1+t\log^2(t))^{-1} \sum_{s=1}^\infty \exp\left(-s\frac{\epsilon^2}{2\sigma_P^2(\mu-\epsilon,\mu)}\right) && \text{(EC.63)} \\
&= (1+t\log^2(t))^{-1}\left(\exp\left(\frac{\epsilon^2}{2\sigma_P^2(\mu-\epsilon,\mu)}\right) - 1\right)^{-1}.
\end{aligned}$$

Note that (EC.60) follows from Lemma EC.4, (EC.61) follows from Lemma EC.3, (EC.62) follows from a Chernoff bound. For sufficiently large t, say $t \geq L$ for some $L > 0$, we have $f(t) \geq \log(1 + t\log^2(t))$, which then yields (EC.63). Then,

$$\begin{aligned}\mathbb{E}[\xi] &= \int_0^\infty \mathbb{P}(\xi > t)dt \\ &\leq L + \left(\exp\left(\frac{\epsilon^2}{2\sigma_P^2(\mu-\epsilon,\mu)}\right) - 1\right)^{-1} \int_L^\infty (1 + t\log^2(t))^{-1} dt \qquad \text{(EC.64)} \\ &=: C(\mu,\epsilon),\end{aligned}$$

where the integral on the right side of (EC.64) is finite, and thus $C(\mu,\epsilon)$ (defined to be equal to the right side of (EC.64)) is a finite constant depending on μ and ϵ. □

LEMMA EC.2 **(Lemma 10.8 from Lattimore and Szepesvári (2020))**. *Let $X_1, \ldots, X_T$ be independent random variables from P^μ, and let $\overline{\mu}_s = s^{-1}\sum_{j=1}^s X_j$. Let $\Delta > 0$ and define*

$$\kappa = \sum_{s=1}^T \mathbb{I}\left(d_P(\overline{\mu}_s, \mu + \Delta) \leq \frac{f(T)}{s}\right). \qquad \text{(EC.65)}$$

Then, for any $\epsilon \in (0, \Delta)$,

$$\mathbb{E}[\kappa] \leq \frac{f(T)}{d_P(\mu+\epsilon, \mu+\Delta)} + C'(\mu,\epsilon), \qquad \text{(EC.66)}$$

where $C'(\mu,\epsilon)$ is a finite constant depending on μ and ϵ.

Proof of Lemma EC.2. Let $\epsilon \in (0,\Delta)$ and $a = f(T)/d_P(\mu+\epsilon,\mu+\Delta)$. Then,

$$\begin{aligned}\mathbb{E}[\kappa] &= \sum_{s=1}^T \mathbb{P}\left(d_P(\overline{\mu}_s, \mu+\Delta) \leq \frac{f(T)}{s}\right) \\ &\leq \sum_{s=1}^T \mathbb{P}\left(\overline{\mu}_s \geq \mu+\epsilon \text{ or } d_P(\mu+\epsilon,\mu+\Delta) \leq \frac{f(T)}{s}\right) \\ &\leq a + \sum_{s=\lceil a\rceil}^T \mathbb{P}(\overline{\mu}_s \geq \mu+\epsilon) \\ &\leq a + \sum_{s=1}^T \exp(-s d_P(\mu+\epsilon,\mu)) \qquad \text{(EC.67)} \\ &\leq a + \sum_{s=1}^\infty \exp\left(-s\frac{\epsilon^2}{2\sigma_P^2(\mu+\epsilon,\mu)}\right), \qquad \text{(EC.68)} \\ &= a + \left(\exp\left(\frac{\epsilon^2}{2\sigma_P^2(\mu+\epsilon,\mu)}\right) - 1\right)^{-1},\end{aligned}$$

where (EC.67) follows from a Chernoff bound, and (EC.68) follows from Lemma EC.3. Then, defining $C'(\mu,\epsilon) := \left(\exp(\epsilon^2/(2\sigma_P^2(\mu+\epsilon,\mu))) - 1\right)^{-1}$, we see that $C'(\mu,\epsilon)$ is a finite constant depending on μ and ϵ. □

LEMMA EC.3 **(Lemma 2 from Appendix of Cappé et al. (2013))**. *For a base distribution P, let $\mu_1, \mu_2 \in \mathcal{I}_P$. Then,*

$$d_P(\mu_1, \mu_2) \geq \frac{(\mu_1 - \mu_2)^2}{2\sigma_P^2(\mu_1, \mu_2)}, \tag{EC.69}$$

where $\sigma_P^2(\mu_1, \mu_2) = \max_{z \in [l,u]} \mathrm{Var}_{X \sim P^z}(X)$, with $l = \min(\mu_1, \mu_2)$ and $u = \max(\mu_1, \mu_2)$.

Proof of Lemma EC.3. See pages 7-8 of the Appendix of Cappé et al. (2013). □

LEMMA EC.4 **(Triangle Inequality for Information Projection)**. *For a base distribution P, let $\epsilon > 0$ and $\mu_1, \mu_2 \in \mathcal{I}_P$ such that $\mu_1 \leq \mu_2 - \epsilon < \mu_2$. Then,*

$$d_P(\mu_1, \mu_2 - \epsilon) + d_P(\mu_2 - \epsilon, \mu_2) \leq d_P(\mu_1, \mu_2). \tag{EC.70}$$

Proof of Lemma EC.4. Let $\mathcal{E} = \{Q : \mu(Q) \leq \mu_2 - \epsilon\}$, the set of all distributions with mean $\leq \mu_2 - \epsilon$, which is a closed convex set of distributions. By Theorem 11.6.1 (on page 367) of Cover and Thomas (2006) (alternatively, see Theorem 15.10 (on page 303) of Polanskiy and Wu (2023)), for $Q^* = \arg\min_{Q \in \mathcal{E}} D(Q \,\|\, P^{\mu_2})$, we have

$$D(P^{\mu_1} \,\|\, Q^*) + D(Q^* \,\|\, P^{\mu_2}) \leq d_P(\mu_1, \mu_2). \tag{EC.71}$$

Moreover, Theorem 15.11 (on pages 303-304) of Polanskiy and Wu (2023) indicates that in this case, $Q^* = P^{\mu_2 - \epsilon}$. So, together with (EC.71), the desired result in (EC.70) is established. □

Appendix EC.6: Proofs for Sections 6.2-6.3

We first establish that distributions in $\mathcal{M}_{P,b}$ obey a Chernoff bound using the re-scaled divergence $d_P/(1+b)$ instead of their usual KL divergence. Recall that the enlarged family of distributions is:

$$\mathcal{M}_{P,b} = \{Q : \mu(Q) \in \mathcal{I}_P,\ \Lambda_Q(\theta) \leq \Psi_{P,b}(\mu(Q), \theta)\ \forall\, \theta \in \mathbb{R}\}, \tag{EC.72}$$

where for any distribution Q and $z \in \mathcal{I}_Q$, we define:

$$\Psi_{Q,b}(z, \theta) = \frac{\Lambda_{Q^z}((1+b)\theta)}{1+b}, \quad \theta \in \mathbb{R}. \tag{EC.73}$$

LEMMA EC.5. *Let $Q \in \mathcal{M}_{P,b}$ and $\overline{\mu}_s = s^{-1} \sum_{j=1}^s X_j$, where $X_1, \ldots, X_s$ are independent random variables from Q. Then,*

$$\mathbb{P}(\overline{\mu}_s > z) \leq \exp\left(-s \frac{d_P(z, \mu(Q))}{1+b}\right), \quad \text{for } z > \mu(Q), \tag{EC.74}$$

$$\mathbb{P}(\overline{\mu}_s < z) \leq \exp\left(-s \frac{d_P(z, \mu(Q))}{1+b}\right), \quad \text{for } z < \mu(Q). \tag{EC.75}$$

Proof of Lemma EC.5. The re-scaled divergence $d_P/(1+b)$ is directly related to the re-scaled CGF's in (EC.73). Specifically, $y \mapsto d_P(y,z)/(1+b)$ is the convex conjugate of $\theta \mapsto \Psi_{P,b}(z,\theta)$, as seen via:

$$\sup_{\theta\in\mathbb{R}}\{\theta y - \Psi_{P,b}(z,\theta)\} = \frac{1}{1+b}\cdot\sup_{\theta\in\mathbb{R}}\{\theta y - \Lambda_{P^z}(\theta)\} = \frac{d_P(y,z)}{1+b}. \tag{EC.76}$$

For $Q \in \mathcal{M}_{P,b}$, using (EC.76), we have

$$\begin{aligned} d_Q(y,\mu(Q)) &= \sup_{\theta\in\mathbb{R}}\{\theta y - \Lambda_Q(\theta)\} \\ &\geq \sup_{\theta\in\mathbb{R}}\{\theta y - \Psi_{P,b}(\mu(Q),\theta)\} = \frac{d_P(y,\mu(Q))}{1+b}. \end{aligned}$$

Therefore, for $z > \mu(Q)$

$$\begin{aligned} \mathbb{P}\left(\overline{\mu}_s > z\right) &\leq \exp(-s d_Q(z,\mu(Q))) \qquad \text{(EC.77)} \\ &\leq \exp\left(-s\frac{d_P(z,\mu(Q))}{1+b}\right), \end{aligned}$$

where (EC.77) follows from a Chernoff bound. So, (EC.74) is established, and (EC.75) is established by the same arguments. □

Next, we have the proofs for Corollary 5 and 6.

Proof of Corollary 5. The above proof of (35) in Proposition 7 can be applied to the new setting in Corollary 5. Recall that we now assume that $f(t) \geq (1+b)\log(1+t\log^2(t))$ for sufficiently large t. (In the above proof of (35), the lower bound was $\log(1+t\log^2(t))$, without the $1+b$ factor.) In Lemma EC.1 and Lemma EC.2, in the definitions of ξ and κ in (EC.59) and (EC.65), respectively, the following two options are equivalent upon inspection.

1. Assume that $f(t) \geq (1+b)\log(1+t\log^2(t))$ for sufficiently large t, and keep d_P unmodified.
2. Assume that $f(t) \geq \log(1+t\log^2(t))$ for sufficiently large t, and replace d_P by $d_P/(1+b)$.

Instead of option 1, we use option 2. Then, the proofs of Lemma EC.1 and Lemma EC.2 go through exactly as before, except that now d_P is replaced by $d_P/(1+b)$ in every instance. Since the distributions of the arm rewards now belong to $\mathcal{M}_{P,b}$, the Chernoff bounds used in (EC.62) and (EC.67) now hold with the re-scaled divergence $d_P/(1+b)$, as justified by Lemma EC.5.

Then, as in the above proof of (35) in Proposition 7, we have established the asymptotic upper bound part of (39), i.e., for sub-optimal arm i in environment $\nu \in \mathcal{M}_{P,b}^K$,

$$\limsup_{T\to\infty}\frac{\mathbb{E}_{\nu\pi}\left[N_i(T)\right]}{f(T)} \leq \frac{1+b}{d_P(\mu(Q_i),\mu_*(\nu))}, \tag{EC.78}$$

where $f(t) \geq \log(1+t\log^2(t))$ for sufficiently large t. The matching asymptotic lower bound part

$$\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu\pi}\left[N_i(T)\right]}{f(T)} \geq \frac{1+b}{d_P(\mu(Q_i),\mu_*(\nu))}. \tag{EC.79}$$

is easily obtained from the WLLN derived from the SLLN of Lemma 3 (see Appendix EC.4), together with Markov's inequality. Together, (EC.78) and (EC.79) yield the desired conclusion of Corollary 5. □

Proof of Corollary 6. The proof of Corollary 6 is essentially the same as that of Corollary 5. The only difference is that we now must use a Chernoff bound for additive functionals of finite state space Markov chains. Theorem 1 of Moulos and Anantharam (2019) provides such a Chernoff bound that is convenient for this purpose. (Earlier and more general results can be found in Miller (1961) and Kontoyiannis and Meyn (2003), respectively.) Recall that in (40) from Section 6.3, we defined the following set of transition matrices:

$$\widetilde{\mathcal{M}}_{P,b} = \left\{ H \in \mathcal{S}_{|S|} : \phi_H(\theta) \leq \Psi_{P,b}(\phi'_H(0), \theta) \ \forall\, \theta \in \mathbb{R} \right\}.$$

Using the same arguments as in Lemma EC.5, we can deduce that Markov chains with transition matrices $H \in \widetilde{\mathcal{M}}_{P,b}$ obey a Chernoff bound involving the re-scaled divergence $d_P/(1+b)$. Specifically, with $\overline{\mu}_s = s^{-1}\sum_{j=1}^{s} X_j$ and $X_1, X_2, \ldots$ evolving according to transition matrix H (with any initial distribution), we have

$$\mathbb{P}\left(\overline{\mu}_s > z\right) \leq c_H \exp\left(-s\frac{d_P(z, \phi'_H(0))}{1+b}\right), \quad \text{for } z > \phi'_H(0),$$
$$\mathbb{P}\left(\overline{\mu}_s < z\right) \leq c_H \exp\left(-s\frac{d_P(z, \phi'_H(0))}{1+b}\right), \quad \text{for } z < \phi'_H(0),$$

where $c_H > 0$ is a constant depending only on the transition matrix H (see Theorem 1 and Proposition 1 of Moulos and Anantharam (2019)). The rest of the proof is identical to that of Corollary 5. □