The Fragility of Optimized Bandit Algorithms

Lin Fan
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, linfan@stanford.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 tail exponent of the resulting regret distribution.

Key words: Multi-armed Bandits, Regret Distribution, Limit Theorems, Mis-specification, Robustness
History: Manuscript version – May 24, 2023

Contents

1 Introduction **2**
1.1 Related Work 5

2 Model and Preliminaries **7**
2.1 The Multi-armed Bandit Framework 7
2.2 Optimized Algorithms 8

3 Characterization of the Regret Distribution Tail **9**
3.1 Truncated Cauchy Tails 9
3.2 Key Ideas Behind Theorem 1 and Further Results 13
3.3 Tail Probability Upper Bounds 15
3.4 Generalized Lower Bounds for Expected Regret 17

4 Illustrations of Fragility **21**
4.1 Mis-specified Reward Distribution . . . 22
4.2 Tail Probability Lower Bounds for General Reward Processes . . . 23
4.3 Mis-specified Reward Dependence Structure . . . 25
4.4 Higher Moments . . . 27

5 Improvement of the Regret Distribution Tail **28**
5.1 A Simple Approach to Obtain Lighter Regret Tails . . . 28
5.2 Robustness to Mis-specified Reward Distribution . . . 30
5.3 Robustness to Mis-specified Reward Dependence Structure . . . 31

6 Proofs of Theorems 1 and 2 **33**
6.1 Proof of Theorem 1 . . . 33
6.2 Proof of Theorem 2 . . . 35

7 Numerical Experiments **38**

A Proofs for Section 3.1 **39**

B Proofs for Section 3.2 **45**

C Proofs for Section 3.3 **45**

D Proofs for Section 3.4 **46**

E Proofs for Section 4.1 **47**

F Proofs for Section 4.2 **49**

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 such 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, Corollary 4 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 5 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 4. 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 as to ensure a desired degree of robustness to model mis-specification. The modification is designed to lighten the regret distribution tail to a given exponent, thereby creating a prescribed margin of safety against model mis-specification. As a part of our analysis, we provide a trade-off between the logarithmic rate of exploration and the resulting heaviness of the regret tail. For example, in well-specified settings, if one increases the 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 $\log^a(T) \leq x \leq cT$, for any $a > 1$ and suitable $c > 0$.

The rest of the paper is structured as follows. After discussing related work in Section 1.1, we introduce the setup for the rest of 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 2), 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 2 does not hold); see Theorem 2. In Section 3.4, we provide an alternative, intuitive proof of the generalized Lai-Robbins lower bound for expected regret (Theorem 3) by focusing on the regret tail and using our change-of-measure arguments. Our proof sheds new light on the result and further provides a sharp trade-off between lighter regret tails and larger expected regret (Proposition 6). In Sections 4.1 and 4.3, we show 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 4.2, where we establish general lower bounds for the regret tail of algorithms such as KL-UCB when the rewards come from stochastic processes (Theorem 4). Moreover, we show in Section 4.4 that such optimized algorithms offer no control over the p'th moment of regret for any $p > 1$. In Section 5.1, building upon Section 3.3, we discuss how to design UCB algorithms to achieve any desired exponent of the regret tail uniformly over a general class of bandit environments. We then discuss how lighter regret tails provide protection against mis-specification of the distribution of rewards and the serial dependence structure of rewards in Sections 5.2 and 5.3, respectively. In Section 7, we examine some numerical experiments. We conclude with the proofs of Theorems 1 and 2 in Sections 6.1 and 6.2, 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:

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

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

Recently, Ashutosh et al. (2021) show that for an algorithm to achieve expected regret of logarithmic order across a collection of bandit 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 expected regret could be of polynomial order. Their result provides no information about algorithm behavior for any particular bandit instance, nor does it cover narrower classes of distributions (e.g., Gaussian).

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), Zimin et al. (2014), Szorenyi et al. (2015), Vakili and Zhao (2016), Galichet et al. (2013), 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.

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. 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 (with π_t defined on $([K] \times \mathbb{R})^t \times 2^{[K]}$). Upon selecting the arm $A(t)$, a reward $Y(t)$ from arm $A(t)$ is received as feedback. 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 $X_i(t)$ to denote the reward received when arm i is played for the t-th instance, 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

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

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 $\mathcal{M}_P^K$, which is stated as follows. Let π be (so-called) $\mathcal{M}_P$-*consistent*, satisfying 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 notion of consistency rules out unnatural algorithms which over-specialize and perform very well in particular environments within a class, but very poorly in others.) Then for 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, i.e., the condition in Definition 1 holds.

Definition 1 (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 Distribution 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 4.1) or serial dependence structure (see Section 4.3) of arm rewards can cause optimized algorithms to suffer expected regret that grows 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 4.4). It is furthermore striking that every environment within such classes of bandit environments suffers from these fragility issues, not just some worst case environments within such classes.

Theorem 1 relies in part on the notion of *discrimination equivalence*, as stated in Definition 2 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 2 (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 $-\kappa$ if $\mathbb{P}_{\nu\pi}(R(T) > x) \asymp x^{-\kappa}$ as $T \to \infty$, uniformly in x with $T^a \leq x \leq cT$, for any $0 < a < 1$ and suitable $c > 0$. 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 Section 6.1. 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 $B_\gamma(T) = [T^\gamma, (1-\gamma)T]$ 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 A.

LEMMA 1. *P is discrimination equivalent if and only if*

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

(Alternatively, $\lim_{\theta\to-\infty} \Lambda_P^*(\Lambda_P'(\theta)) = \infty$, 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 A.

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 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 greater the relative difficulty, the closer the ratio is to 1. 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.

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 $P(dx) = e^x \cdot \mathbb{I}(x \leq 0)\, dx$ 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 environments* within a class, Corollary 1 is weaker as it pertains only to *some environments* 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 $0 < \mu_{r(1)} - \mu_{r(2)} < \delta$,*

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

with $B_\gamma(T) = [T^\gamma, (1-\gamma)T]$ 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. Proposition 3 below provides such control through a weak law of large numbers (WLLN) for the number of sub-optimal arm plays of optimized algorithms. Proposition 3 follows immediately for optimized algorithms due to a "one-sided" and more general version of the result in Proposition 5 in Section 3.4.

PROPOSITION 3. *Let π be $\mathcal{M}_P$-optimized. Then for any environment $\nu = (P^{\mu_1}, \dots, P^{\mu_K}) \in \mathcal{M}_P^K$ 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 $c \in (0,1)$, 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:

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

If additionally P is discrimination equivalent, then

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

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 $\widetilde{\nu}$,

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

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 Proposition 3,

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

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 $\widetilde{\nu}$) $\mathbb{P}_{\widetilde{\nu}\pi}(N_2(T) > cT) \to 1$, using (13) and (16), we obtain:

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

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,

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

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 4 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 4 is given in Appendix B. (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 4. *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 $\epsilon > 0$,*

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

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 $i \geq 2$), 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 Section 6.1.

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 $i \geq 2$) 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 $i \geq 2$, 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 Section 6.2.

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 3.4, 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 Proposition 5 and Theorem 3.)

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 5, 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 7, we develop a result for general UCB algorithms operating in essentially arbitrary environments, including mis-specified ones.

3.4. Generalized Lower Bounds for Expected Regret

In Corollary 1, we saw that for any optimized algorithm, there always exist environments with very close top arm means for which the algorithm produces regret tails that are arbitrarily close to being (if not exactly) truncated Cauchy. This suggests that one cannot further reduce the expected regret of an optimized algorithm (by exploring less/exploiting more) or else the regret tails will become heavier than truncated Cauchy in such environments and violate consistency. In this section, we make this heuristic rigorous and develop an alternative proof of a generalized version of the Lai-Robbins lower bound for expected regret. This result, as stated in Theorem 3 below, was first developed in Proposition 1 of Burnetas and Katehakis (1996) (by extending Theorem 2 of Lai and Robbins (1985)). Our proof of Theorem 3, which is essentially contained in Proposition 5, directly mirrors the proof of our main result, Theorem 1 (see Section 6.1), as discussed in Remark 2 below. Furthermore, in Proposition 6, we use this approach to develop an asymptotic lower bound on expected regret, given an upper bound on the regret tail, which establishes a sharp trade-off between the two.

Building on the Lai-Robbins lower bound (where the model $\mathcal{M}_P$ is an exponential family as in (1)-(2)), in Theorem 3, the model denoted by $\mathcal{M}$ is allowed to be an arbitrary collection of distributions (possibly even finitely many) with finite means. For such an arbitrary $\mathcal{M}$, an arbitrary distribution Q, and $z \in \mathbb{R}$, we define:

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

where $D(Q \,\|\, Q')$ denotes the KL divergence between the distributions Q and Q', and we take the infimum of the empty set to be $+\infty$. Moreover, in Theorem 3, the environments in the corresponding environment class $\mathcal{M}^K$ are allowed to be arbitrary. The best arm(s), second-best arm(s), etc., do not need to be unique.

THEOREM 3. *Let the model $\mathcal{M}$ consist of an arbitrary collection of distributions with finite means. Let π be $\mathcal{M}$-consistent, i.e., π satisfies (4) for the general class $\mathcal{M}^K$. Then for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and each sub-optimal arm i,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{\log(T)} \geq \frac{1}{D_{\inf}(P_i, \mu_*(\nu), \mathcal{M})}. \tag{19}$$

Theorem 3 follows immediately from Proposition 5 below. When $\mathcal{M}$ is an exponential family model as in (1)-(2), (19) simplifies to (5). Moreover, when $\mathcal{M}$ is an exponential family, Proposition 5 directly implies Proposition 3. In the proof, 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'.

PROPOSITION 5. *Under the assumptions of Theorem 3, for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and each sub-optimal arm i,*

$$\lim_{T\to\infty} \mathbb{P}_{\nu\pi}\left(\frac{N_i(T)}{\log(T)} \geq \frac{1}{D_{\inf}(P_i, \mu_*(\nu), \mathcal{M})}\right) = 1. \tag{20}$$

Proof of Proposition 5. Suppose there is an environment $\widetilde{\nu} = (\widetilde{P}_1, P_2, \ldots, P_K) \in \mathcal{M}^K$ for which (20) 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

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(\frac{N_1(T_n)}{\log(T_n)} \leq \frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1, \mu(P_2), \mathcal{M})}\right) \geq \epsilon. \tag{21}$$

Denote the event in (21) 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

$$\frac{D(\widetilde{P}_1 \,\|\, P_1)}{D_{\inf}(\widetilde{P}_1, \mu(P_2), \mathcal{M})} \leq 1 + \epsilon. \tag{22}$$

(Such P_1 exists or else $D_{\inf}(\widetilde{P}_1, \mu(P_2), \mathcal{M}) = \infty$ and (20) 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:

$$\mathcal{B}_n = \left\{\left|\frac{1}{N_1(T_n)} \sum_{t=1}^{N_1(T_n)} \log\frac{dP_1}{d\widetilde{P}_1}(X_1(t)) + D(\widetilde{P}_1 \,\|\, P_1)\right| \leq \delta\right\}$$
$$\mathcal{C}_n = \{\exists\, i \neq 1 : N_i(T_n) > T_n/(2K)\}.$$

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

$$\mathbb{P}_{\nu\pi}(\mathcal{C}_n) \geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{C}_n\right) \prod_{t=1}^{N_1(T_n)} \frac{dP_1}{d\widetilde{P}_1}(X_1(t))\right] \tag{23}$$

$$\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_n, \mathcal{B}_n\right) \exp\left(\frac{1}{N_1(T_n)} \sum_{t=1}^{N_1(T_n)} \log \frac{dP_1}{d\widetilde{P}_1}(X_1(t)) \cdot N_1(T_n)\right)\right] \tag{24}$$

$$\geq \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_n, \mathcal{B}_n) \cdot \exp\left(-\left(D(\widetilde{P}_1 \,\|\, P_1) + \delta\right) \cdot \frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1, \mu(P_2), \mathcal{M})} \log(T_n)\right), \tag{25}$$

where (24) follows from $\mathcal{C}_n \supset \mathcal{A}_n$ for large n, and (25) follows from lower bounds using $\mathcal{A}_n$ and $\mathcal{B}_n$. By Lemma 2 (in Appendix D) and the WLLN for sample means, $\lim_{n\to\infty} \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_n) = 1$. So from (21), $\liminf_{n\to\infty} \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_n, \mathcal{B}_n) \geq \epsilon$. From (25), taking logs and dividing by $\log(T_n)$, sending $n \to \infty$ followed by $\delta \downarrow 0$, and then applying (22), we obtain:

$$\liminf_{n\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(\mathcal{C}_n)}{\log(T_n)} \geq \liminf_{n\to\infty} \frac{\log \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_n, \mathcal{B}_n)}{\log(T_n)} - (1-\epsilon)\frac{D(\widetilde{P}_1 \,\|\, P_1)}{D_{\inf}(\widetilde{P}_1, \mu(P_2), \mathcal{M})} \geq -(1-\epsilon^2). \tag{26}$$

Since $\epsilon \in (0,1)$, this violates the $\mathcal{M}$-consistency of π, and thus (21) cannot be true. □

Remark 2. As mentioned at the beginning of the current section, the proof of Proposition 5 mirrors that of the proof of Theorem 1 (in Section 6.1). Specializing Proposition 5 so that $\mathcal{M} = \mathcal{M}_P$ (an exponential family model), we would have $\widetilde{\nu} = (P^{\widetilde{\mu}_1}, P^{\mu_2}, \ldots, P^{\mu_K})$ and $\nu = (P^{\mu_1}, P^{\mu_2}, \ldots, P^{\mu_K})$, where μ_1 and μ_2 are the two highest means, with $\mu_1 > \mu_2 > \widetilde{\mu}_1$. Moreover, throughout the proof, $D(\widetilde{P}_1 \,\|\, P_1)$ would be replaced by $d_P(\widetilde{\mu}_1, \mu_1)$, and $D_{\inf}(\widetilde{P}_1, \mu(P_2), \mathcal{M})$ by $d_P(\widetilde{\mu}_1, \mu_2)$. Then, the steps (23)-(26) mirror the steps (51)-(55) in the proof of Theorem 1. As mentioned previously, for such an environment ν with arbitrarily close top arm means μ_1 and μ_2, Theorem 1 and Corollary 1 indicate that the regret tail of an $\mathcal{M}_P$-optimized algorithm is arbitrarily close to being truncated Cauchy, with exponent $\geq -d_P(\widetilde{\mu}_1, \mu_1)/d_P(\widetilde{\mu}_1, \mu_2)$. So, for an $\mathcal{M}_P$-consistent algorithm π, one cannot further reduce the expected regret in environment $\widetilde{\nu}$ beyond that of an optimized algorithm by allowing (21) to hold, as it would cause the regret tail to be heavier than truncated Cauchy in such an environment ν, as can be seen from (26).

Remark 3. To the best of our knowledge, there are three alternative proof techniques relevant to Theorem 3 that exist in the current literature: Proposition 1 of Burnetas and Katehakis (1996), Theorem 1 of Garivier et al. (2019) and Theorem 16.2 of Lattimore and Szepesvári (2020). Compared to the proof of Burnetas and Katehakis (1996), our proof focuses on the heaviness of the regret tail, with direct connections to our regret tail characterizations for optimized algorithms (see Remark 2 above). The differences between our proof and those of Garivier et al. (2019) and Lattimore and Szepesvári (2020) are quite pronounced. The proofs in these two works utilize

general-purpose information-theoretic results, while we argue directly using change of measure and lower-bounding the resulting likelihood ratio. Additionally, Kaufmann et al. (2016) (in Theorem 21) develop a similar asymptotic lower bound for expected regret using tools tailored for best-arm identification problems. However, unlike Theorem 3 here and the results of Burnetas and Katehakis (1996), Garivier et al. (2019) and Lattimore and Szepesvári (2020), the result in Kaufmann et al. (2016) requires the optimal arm to be unique.

By slightly modifying the proof of Proposition 5, we obtain a generalization of that result and of Theorem 3. The generalization, as stated in Proposition 6 below, establishes trade-off between a lighter regret tail and a modest increase in expected regret. Here, the standard assumption of consistency is replaced by the upper bound on the regret tail in (27). (For the case $b = 0$, $\mathcal{M}$-consistency implies (27).) We will refer to Proposition 6 in Section 5 when we consider modifications of KL-UCB to obtain lighter regret tails.

PROPOSITION 6. *Let the model $\mathcal{M}$ consist of an arbitrary collection of distributions with finite means, and let $b \geq 0$. For every environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$, suppose π satisfies for each sub-optimal arm i and any $\gamma \in (0,1)$:*

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_i(T) > \gamma T)}{\log(T)} \leq -(1+b). \tag{27}$$

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

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{\log(T)} \geq \frac{1+b}{D_{\inf}(P_i, \mu_*(\nu), \mathcal{M})}. \tag{28}$$

Proof of Proposition 6. We use almost the same proof of Proposition 5 to show that for any $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and each sub-optimal arm i,

$$\lim_{T\to\infty} \mathbb{P}_{\nu\pi}\left(\frac{N_i(T)}{\log(T)} \geq \frac{1+b}{D_{\inf}(P_i, \mu_*(\nu), \mathcal{M})}\right) = 1. \tag{29}$$

The only difference is that we modify (21) to be:

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(\frac{N_1(T_n)}{\log(T_n)} \leq \frac{(1-\epsilon)(1+b)}{D_{\inf}(\widetilde{P}_1, \mu(P_2), \mathcal{M})}\right) \geq \epsilon. \tag{30}$$

(In Lemma 2, the $\mathcal{M}$-consistency assumption can be replaced by (27) with any $b \geq 0$ to yield the same conclusion.) The modification in (30) changes (25)-(26) accordingly, and leads to violation of (27), thus establishing (29). Then, (28) follows from (29) and Markov's inequality. □

4. Illustrations of Fragility

In this section, we highlight several ways in which optimized algorithms are fragile. To do so, our main focus will be on the development of regret tail characterizations for bandit algorithms in 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 often some degree of mis-specification. So it is important to have some understanding of how vulnerable an algorithm is to different forms of mis-specification.

In Section 4.1, we consider mis-specification of the marginal distributions of rewards in iid settings. In Section 4.2, we develop lower bounds on the regret tail for general reward processes, which are then applied to study mis-specification of the serial dependence structure of rewards in Section 4.3. Our analysis of mis-specification in these sections involves stylized departures from the model assumptions built into an algorithm's design. For example, for an algorithm optimized for environments yielding iid Gaussian rewards with a specified variance, we consider what happens to the regret tail when the rewards are iid Gaussian, but with a variance larger than that specified in the algorithm. In another direction, we consider what happens to the regret tail of the same algorithm when the marginal distributions of the rewards are Gaussian with the correct variance, but the rewards are not independent and instead evolve as AR(1) processes. Our analysis, though stylized, reveals that optimized algorithms are highly fragile. The slightest degree of mis-specification, of which there are many forms, can result in regret tails that are heavier than truncated Cauchy, and thus preclude logarithmic expected regret.

To illustrate how the regret tail behaves under model mis-specification, we will focus primarily on the KL-UCB algorithm throughout Sections 4.1-4.3. Unless specified otherwise, our theory will 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, we will always 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. Of course, this ensures that the KL divergence function d_P is always well-defined when sample means of arm rewards are used as the arguments. For simplicity, in Section 4, we will always assume that the true environment ν has a unique optimal arm.

In Section 4.4, we conclude our illustrations of fragility by examining 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).

4.1. 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 7, 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 (31) in Proposition 7 to the right side of (18) in Theorem 2. In well-specified settings, Theorem 2 and Proposition 7 are the same result. In mis-specified settings, which is covered by Proposition 7, 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 7 is given in Appendix E. The proof uses a LLN (Proposition 8) for the regret of (possibly) mis-specified KL-UCB, and a general tail probability lower bound (Theorem 4), which are deferred to Section 4.2. These supporting results are developed for more general (possibly non-iid) reward processes. They are useful for establishing the results in Section 4.3, but they are stronger than needed in the current section.

PROPOSITION 7. *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 $r(i)$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{z\in\mathcal{I}_{Q_{r(j)}} : z<\mu(Q_{r(i)})} \frac{d_{Q_{r(j)}}(z, \mu(Q_{r(j)}))}{d_P(z, \mu(Q_{r(i)}))} \tag{31}$$

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 E. See Figure 1 in Section 7 for numerical simulations illustrating (32).

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 $\sigma_0^2 > 0$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(2)}(T) > x)}{\log(x)} = -\frac{\sigma^2}{\sigma_0^2} \tag{32}$$

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).

4.2. Tail Probability Lower Bounds for General Reward Processes

In this section, we develop supporting results, which are needed in Section 4.3 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 4.1 in settings where the marginal reward distributions are mis-specified. Proposition 8 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 3 below. In Theorem 4, 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,

$$\frac{1}{n}\sum_{t=1}^{n} X_i(t) \to \overline{\Lambda}_i'(0) \tag{33}$$

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 4.3, 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 Proposition 8 and Theorem 4, we introduce the following notion, which can be compared to the notion of an $\mathcal{M}_P$-optimized algorithm in Definition 1.

DEFINITION 3 (PATHWISE CONVERGENT ALGORITHM). An algorithm π is *$\mathcal{M}_P$-pathwise convergent* if for any environment ν yielding arm reward sequences $X_i(t)$, $i \in [K]$, $t \geq 1$,

$$\left\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} X_i(t) = c_i,\ i \in [K]\right\} \subset \left\{\omega : \lim_{T\to\infty} \frac{N_i(T)}{\log(T)} = \frac{1}{d_P(c_i, \max_j c_j)},\ \forall\, i \neq \arg\max_j c_j\right\}. \quad (34)$$

We conjecture that, in general, $\mathcal{M}_P$-optimized algorithms are also $\mathcal{M}_P$-pathwise convergent. This is directly supported by Proposition 8 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 Proposition 8 is based on the arguments in Cowan and Katehakis (2019), and can be found in Appendix F.

PROPOSITION 8. *$\mathcal{M}_P$-optimized KL-UCB is $\mathcal{M}_P$-pathwise convergent.*

We now introduce Theorem 4, whose proof can be found in Appendix F. 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 4. *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 $r(i)$,*

$$\liminf_{T\to\infty} \inf_{x \in B_\gamma(T)} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} \geq -\sum_{j=1}^{i-1} \inf_{z \in \overline{\mathcal{I}}_{r(j)} : z < \overline{\Lambda}'_{r(i)}(0)} \frac{\overline{\Lambda}^*_{r(j)}(z)}{d_P(z, \overline{\Lambda}'_{r(i)}(0))}, \quad (35)$$

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

REMARK 4. Whenever we can establish a WLLN for the $N_i(T)$ (e.g., as in Proposition 3), 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 4). The almost sure convergence of the $N_i(T)$, as provided by Assumptions 1-2 (leading to (33)) together with pathwise convergence (in Definition 3), is sufficient but not necessary.

4.3. 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:

$$X_i(t) = \alpha_i + \beta_i X_i(t-1) + W_i(t), \tag{36}$$

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 (33)) 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 4.

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 (36)),*

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

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,

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(2)}(T) > T/2)}{\log(T)} \geq -\frac{1-\beta_0}{1+\beta_0}, \tag{37}$$

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 (37) through numerical simulations in Figure 2 in Section 7. The simulations suggest that the lower bound in (37) is tight.

In Corollary 4 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:

$$\big(\exp(\theta \cdot y) H(x,y),\ x,y \in S\big). \tag{38}$$

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 Corollary 4 as it does in Theorem 4.) 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 (33)) 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 (38). 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 Corollary 4 follows from a straightforward verification of Assumptions 1-2, which is omitted, and then a direct application of Theorem 4. To establish the upper bound part, we can again use the proof of Theorem 2 (in Section 6.2) and substitute in, where appropriate (in (63) and (68)), 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.)

COROLLARY 4. *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 $r(i)$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{z \in \overline{\mathcal{I}}_{r(j)} : z < \overline{\Lambda}'_{r(i)}(0)} \frac{\overline{\Lambda}^*_{r(j)}(z)}{d_P(z, \overline{\Lambda}'_{r(i)}(0))} \tag{39}$$

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 (39). 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:

$$H_{r(1)} = \begin{bmatrix} 1-q & q \\ 1-w(q) & w(q) \end{bmatrix}. \tag{40}$$

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 (39) 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 (39) 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 (39). Here, $i = 2$, and we consider different values of q (as used in the best arm's transition matrix $H_{r(1)}$ in (40)) and Δ (the gap between the equilibrium means of the two best arms, $r(1)$ and $r(2)$).

4.4. 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 Proposition 3, the $1 + \delta$ moment of regret is not poly-logarithmic. In Corollary 5 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 5. *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. Improvement of the Regret Distribution Tail

In Sections 3 and 4, we have seen how optimized algorithms prioritize expected regret minimization at the cost of rendering the tail of the regret distribution susceptible to even small degrees of model mis-specification. In Section 5, we discuss a general approach to make the regret tail lighter (with a more negative tail exponent), which as we show, leads to a degree of robustness to model mis-specification. Specifically, in Section 5.1, we describe a simple way to construct a UCB algorithm so that the regret tail exponent is $-(1+b)$ (or less) for any desired $b>0$, for an exponential family class of environments. We then show how this also makes the regret tail suitably lighter uniformly over a general class of environments. In Section 5.2, we show that the modification provides protection against mis-specification of the arm reward distributions in iid settings. In Section 5.3, we show that such modification also provides protection against Markovian departures from independence of the arm rewards. Our analysis further establishes an explicit trade-off between the amount of UCB exploration (and expected regret) and the resulting heaviness of the regret tail (see Remarks 6 and 7).

5.1. A Simple Approach to Obtain Lighter Regret Tails

In this section, and in Sections 5.2 and 5.3, we focus on Algorithm 1, which is a simple modification of the KL-UCB algorithm (Algorithm 2 of Cappé et al. (2013)). Like KL-UCB, Algorithm 1 is defined for any exponential family $\mathcal{M}_P$, as in (1)-(2). However, the difference is that Algorithm 1 involves re-scaling the KL divergence d_P by $1/(1+b)$, for a desired $b>0$. This has the effect of inducing additional exploration, and is equivalent to increasing the "radius" of the upper confidence bound by a factor of $1+b$; see also Remark 6 towards the end of this section.

Algorithm 1 : b-robustified KL-UCB (based on Algorithm 2 of Cappé et al. (2013))

input: $b \geq 0$, $d_P : \mathcal{I}_P \times \mathcal{I}_P \to [0,\infty)$

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

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

Play the arm (with ties broken arbitrarily):

$$A(t+1) = \arg\max_{i\in[K]} \sup\left\{ z \in \mathcal{I}_P : \frac{d_P(\widehat{\mu}_i(t), z)}{1+b} \leq \frac{\log(1+t\log^2(t))}{N_i(t)} \right\}$$

end for

Using Algorithm 1, we can ensure that the regret tail exponent is $-(1+b)$ (or less) for all environments in $\mathcal{M}_P^K$. This follows from a direct adaptation of Proposition 7. Specifically, with π as Algorithm 1, we have for any environment $\nu = (P^{\mu_1}, \dots, P^{\mu_K}) \in \mathcal{M}_P^K$ and the i-th-best arm $r(i)$,

$$\begin{aligned} \lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} &= -(1+b) \sum_{j=1}^{i-1} \inf_{z \in \mathcal{I}_P \,:\, z < \mu_{r(i)}} \frac{d_P(z, \mu_{r(j)})}{d_P(z, \mu_{r(i)})} \\ &\le -(1+b)(i-1) \end{aligned} \tag{41}$$

uniformly for $x \in [\log^{1+\gamma}(T), (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$. The numerators in the infima of (41) can be compared to those of (31) in Proposition 7. In Figure 4 in Section 7, we provide numerical simulations that illustrate the result (41) when $\mathcal{M}_P$ is a Gaussian family.

Furthermore, using Algorithm 1, we can ensure that the regret tail exponent is $-(1+b)$ (or less) for all environments in a class $\mathcal{M}_{P,0}^K$ that is larger than $\mathcal{M}_P^K$. This is established in Corollary 6 below. Here, $\mathcal{M}_{P,0}$ is a general family of distributions, whose CGF's are dominated by those of $\mathcal{M}_P$:

$$\mathcal{M}_{P,0} = \{Q : \mu(Q) \in \mathcal{I}_P,\ \Lambda_Q(\theta) \le \Lambda_{P^{\mu(Q)}}(\theta)\ \forall\, \theta \in \mathbb{R}\}\,. \tag{42}$$

(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).) We have the following examples of $\mathcal{M}_P$ and $\mathcal{M}_{P,0}$.

Example 6. 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 7. Let $\mathcal{M}_P$ be the Bernoulli family. Then $\mathcal{M}_{P,0}$ is the family of all distributions supported on a subset of $[0,1]$.

Corollary 6 follows from a direct adaptation of the proof of Proposition 7, similar to the justification for (41). This is because the distributions in $\mathcal{M}_{P,0}$ obey the same Chernoff bounds as those in $\mathcal{M}_P$, per the definition in (42).

Corollary 6. *Let π be Algorithm 1, with divergence d_P and $b \ge 0$. Then for any environment $\nu = (Q_1, \dots, Q_K) \in \mathcal{M}_{P,0}^K$ and the i-th-best arm $r(i)$,*

$$\begin{aligned} \lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{r(i)}(T) > x)}{\log(x)} &= -(1+b) \sum_{j=1}^{i-1} \inf_{z \in \mathcal{I}_{Q_{r(j)}} \,:\, z < \mu(Q_{r(i)})} \frac{d_{Q_{r(j)}}(z, \mu(Q_{r(j)}))}{d_P(z, \mu(Q_{r(i)}))} \\ &\le -(1+b)(i-1) \end{aligned} \tag{43}$$

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

REMARK 5. The $-(1+b)(i-1)$ upper bound on (43) is due to the fact that the ratios in the infima of (43) are at least 1. This is because for any $Q \in \mathcal{M}_{P,0}$,

$$\begin{aligned} d_Q(z, \mu(Q)) &= \sup_{\theta \in \mathbb{R}} \{\theta z - \Lambda_Q(\theta)\} \\ &\geq \sup_{\theta \in \mathbb{R}} \{\theta z - \Lambda_{P^{\mu(Q)}}(\theta)\} = d_P(z, \mu(Q)), \end{aligned}$$

where the inequality follows from the definition of $\mathcal{M}_{P,0}$ in (42).

REMARK 6. From (41) and (43), we see there is an explicit trade-off between the amount of exploration and the resulting heaviness of the regret distribution tail. Specifically, using the re-scaled divergence function $d_P/(1+b)$ instead of d_P in Algorithm 1 is equivalent to increasing the amount of UCB exploration by $1+b$ times, which for a fixed instance of bandit environment ν, yields a regret tail exponent of $-C(\nu) \cdot (1+b)$, where $C(\nu) \geq 1$ is a constant depending on ν. 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 (41) and (43). 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 7. As expected from Proposition 6, aiming for a lighter regret tail does come at the cost of greater expected regret. However, the expected regret increase is modest, being only a multiple of $\log(T)$. And one benefit is greater robustness to model mis-specification, as we will see in the next two sections. (As a follow-up to Corollary 6, we have a precise characterization of the expected regret growth of Algorithm 1 in (48) of Corollary 7 in the next section.)

5.2. 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 as a power of the time horizon T. As we saw in Section 4.1 via Proposition 7 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. We will see in Corollary 7 below that this provides protection against distributional mis-specification of the arm rewards. In particular, we can maintain logarithmic expected regret for environments from an enlarged class $\mathcal{M}^K_{P,b}$, which is defined in (44) below and depends on the chosen value of $b > 0$ in Algorithm 1.

The enlarged family of distributions is:

$$\mathcal{M}_{P,b} = \{Q : \mu(Q) \in \mathcal{I}_P,\ \Lambda_Q(\theta) \leq \Psi_{P,b}(\mu(Q), \theta)\ \forall\, \theta \in \mathbb{R}\}, \tag{44}$$

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

$$\Psi_{Q,b}(z,\theta) = \frac{\Lambda_{Q^z}((1+b)\theta)}{1+b}, \quad \theta \in \mathbb{R}. \tag{45}$$

Setting $b=0$ recovers $\mathcal{M}_{P,0}$ as in (42). Using Jensen's inequality and the definition in (45), it is straightforward to see that $\mathcal{M}_P \subsetneqq \mathcal{M}_{P,b}$ for $b>0$. Moreover, with $\mathcal{M}_{P,0}$ from (42) 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 8. 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 6.)

REMARK 8. Algorithm 1 uses the divergence d_P re-scaled by $1/(1+b)$. This is directly related to the re-scaled CGF's in (45), namely $y \mapsto d_P(y,z)/(1+b)$ is the convex conjugate of $\theta \mapsto \Psi_{P,b}(z,\theta)$, as seen via:

$$\frac{d_P(y,z)}{1+b} = \frac{1}{1+b} \cdot \sup_{\theta\in\mathbb{R}} \{\theta y - \Lambda_{P^z}(\theta)\} = \sup_{\theta\in\mathbb{R}} \{\theta y - \Psi_{P,b}(z,\theta)\}. \tag{46}$$

In Corollary 7, (47) can be obtained by using a straightforward adaptation of the proof of Theorem 10.6 in book by Lattimore and Szepesvári (2020). The proof there is developed for KL-UCB in the iid Bernoulli rewards setting. However, the proof can be directly extended to cover KL-UCB for any exponential family by using a Chernoff bound based on the KL divergence for that family. Moreover, the proof can be directly adapted to our setting in Corollary 7 using $d_P/(1+b)$ as the divergence in the Chernoff bound. Due to (44)-(46), the distributions in $\mathcal{M}_{P,b}$ obey a Chernoff bound that involves the re-scaled divergence $d_P/(1+b)$. Then, (48) follows from Corollary 6 and Proposition 6.

COROLLARY 7. *Let π be Algorithm 1, with divergence d_P and $b \geq 0$. Then for any environment $\nu = (Q_1,\ldots,Q_K) \in \mathcal{M}_{P,b}^K$ and each sub-optimal arm i,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{\log(T)} \leq \frac{1+b}{d(\mu(Q_i),\mu_*(\nu))}. \tag{47}$$

Moreover, for any environment $\nu = (Q_1,\ldots,Q_K) \in \mathcal{M}_{P,0}^K$ and each sub-optimal arm i,

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_i(T)]}{\log(T)} = \frac{1+b}{d(\mu(Q_i),\mu_*(\nu))}. \tag{48}$$

5.3. 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 as a power of the time horizon T. We saw an example of this in Section 4.3 via Corollary 4, particularly via Example 5. To alleviate such issues, we use Algorithm 1. We will see in Corollary 8 below that this provides protection against Markovian departures from independence of the arm rewards. In particular, we can maintain 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 (49) below and depends on the chosen value of $b \geq 0$ in Algorithm 1.

Let $\mathcal{S}_{|S|}$ denote the set of $|S| \times |S|$ irreducible stochastic matrices satisfying the Doeblin Condition (as discussed in Section 4.3 in the context of Corollary 4). We define

$$\widetilde{\mathcal{M}}_{P,b} = \left\{ H \in \mathcal{S}_{|S|} \,:\, \phi_H(\theta) \leq \Psi_{P,b}(\phi'_H(0), \theta) \;\forall\, \theta \in \mathbb{R} \right\}, \tag{49}$$

and we recall that $\phi_H(\theta)$ is the logarithm of the Perron-Frobenius eigenvalue of the tilted version (as in (38)) 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 9, which is given after Corollary 8, 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.

We have Corollary 8 below, which (like Corollary 7) can also be obtained by using a straightforward adaptation of the proof of Theorem 10.6 in Lattimore and Szepesvári (2020) (using $d_P/(1+b)$ as the divergence). Theorem 1 of Moulos and Anantharam (2019) provides a Chernoff bound for additive functionals of finite state space Markov chains that is convenient for this purpose. (As mentioned previously, earlier and more general results can be found in Miller (1961) and Kontoyiannis and Meyn (2003), respectively.) Due to (49) and (45)-(46), Markov chains with transition matrices in $\widetilde{\mathcal{M}}_{P,b}$ obey this Chernoff bound, which involves the re-scaled divergence $d_P/(1+b)$.

Corollary 8. *Let π be Algorithm 1, with divergence d_P and $b \geq 0$. 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,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}\left[N_i(T)\right]}{\log(T)} \leq \frac{1+b}{d_P(\phi'_{H_i}(0), \phi'_{H_{r(1)}}(0))}.$$

Example 9. 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:

$$H = \begin{bmatrix} 1-q & q \\ 1-q' & q' \end{bmatrix}. \tag{50}$$

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 8 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 (50) belongs in $\widetilde{\mathcal{M}}_{P,b}$, as defined in (49) (and (45)).

REMARK 9. 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 to complement the lower bound in Theorem 4 (unlike in Proposition 7 and Corollary 4). 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 8, 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. (The lower bound in Theorem 4 seems to be tight in greater generality than what we are able to provably show.)

6. Proofs of Theorems 1 and 2

6.1. 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 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 $\widetilde{\nu}$,

$$\mathbb{P}_{\nu\pi}(N_2(T) > (1-\gamma)T) = \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(N_2(T) > (1-\gamma)T\right) \prod_{t=1}^{N_1(T)} \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t))\right] \tag{51}$$

$$\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(\frac{1}{N_1(T)} \sum_{t=1}^{N_1(T)} \log \frac{dP^{\mu_1}}{dP^{\widetilde{\mu}_1}}(X_1(t)) \cdot N_1(T)\right)\right] \tag{52}$$

$$\geq \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T, \mathcal{B}_T) \cdot \exp\left(-\left(d_P(\widetilde{\mu}_1, \mu_1) + \delta\right)\left(\frac{1}{d_P(\widetilde{\mu}_1, \mu_2)} + \delta\right) \log(T)\right). \tag{53}$$

where (52) follows from $\{N_2(T) > (1-\gamma)T\} \supset \mathcal{A}_T$ for sufficiently large T, and (53) follows from lower bounds using $\mathcal{A}_T$ and $\mathcal{B}_T$. From (53), taking logs and dividing by $\log(T)$,

$$\frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > (1-\gamma)T)}{\log(T)} \geq \frac{\log \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T, \mathcal{B}_T)}{\log(T)} - \left(d_P(\widetilde{\mu}_1, \mu_1) + \delta\right)\left(\frac{1}{d_P(\widetilde{\mu}_1, \mu_2)} + \delta\right). \tag{54}$$

Using Proposition 3 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 (54) is negligible as $T \to \infty$, and upon sending $\delta \downarrow 0$ and optimizing with respect to $\widetilde{\mu}_1$, we have

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > (1-\gamma)T)}{\log(T)} \geq - \inf_{\widetilde{\mu}_1 \in \mathcal{I}_P : \widetilde{\mu}_1 < \mu_2} \frac{d_P(\widetilde{\mu}_1, \mu_1)}{d_P(\widetilde{\mu}_1, \mu_2)}. \tag{55}$$

The conclusion (7) holds with the infimum over $B_\gamma(T) = [T^\gamma, (1-\gamma)T]$ due to $x \mapsto \log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ being monotone decreasing for $x \in B_\gamma(T)$, with T fixed.

We now establish (8). Because P is discrimination equivalent, the right side of (55) is equal to -1. Because π is $\mathcal{M}_P$-optimized, using Markov's inequality, the case $x = (1-\gamma)T$ in (8) is established, i.e.,

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > (1-\gamma)T)}{\log((1-\gamma)T)} = -1. \tag{56}$$

To obtain the uniform result for $x \in [T^\gamma, (1-\gamma)T]$, note that for $T > T^\gamma/(1-\gamma)$,

$$\mathbb{P}_{\nu\pi}(N_2(T) > T^\gamma) \geq \mathbb{P}_{\nu\pi}(N_2(\lceil T^\gamma/(1-\gamma)\rceil) > T^\gamma). \tag{57}$$

Using (56), but with $\lceil T^\gamma/(1-\gamma)\rceil$ in the place of T, together with Markov's inequality (with π being $\mathcal{M}_P$-optimized),

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(\lceil T^\gamma/(1-\gamma)\rceil) > T^\gamma)}{\log(T^\gamma)} = -1.$$

Thus, using (57) and Markov's inequality (with π being $\mathcal{M}_P$-optimized), the case $x = T^\gamma$ in (8) is established, i.e.,

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > T^\gamma)}{\log(T^\gamma)} = -1. \tag{58}$$

Since, for each T, $x \mapsto \log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ is a monotone decreasing function for $x > 1$, the desired uniform convergence in (8) for $x \in [T^\gamma, (1-\gamma)T]$ follows from the matching limits at the endpoints $x = (1-\gamma)T$ and $x = T^\gamma$, as established in (56) and (58), respectively.

We now show (7) for 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 $\widetilde{\nu}$.) The events $\mathcal{A}_T$ and $\mathcal{B}_T$ become:

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

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

To obtain (7) for sub-optimal arm $i \geq 3$, we can then run through arguments analogous to those in (51)-(55). 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. □

6.2. Proof of Theorem 2

Without loss of generality, suppose that $\mu_1 > \mu_2 > \cdots > \mu_K$ (i.e., $r(i) = i$ for all $i \in [K]$). Define for arm i the KL-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 $\tau_i(n)$ denotes the time of the n-th play of arm i, and as defined previously, $\widehat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{s=1}^{N_i(t)} X_i(s)$. Here, we use the choice $f(t) = \log(t)$, where (as in Algorithm 2 of Cappé et al. (2013)) the "exploration function" $f(t)$ is a design choice. Section 7 of Cappé et al. (2013) recommends using this particular choice of $f(t)$. Moreover, our proof below can easily accommodate other choices such as $f(t) = \log(t) + 3\log\log(t)$ (used in Theorem 1 of Cappé et al. (2013)), or $f(t) = \log(1 + t\log^2(t))$ (used in Theorem 10.6 of Lattimore and Szepesvári (2020)). Using any one of these variations of $f(t)$ does not affect the conclusion of our Theorem 2.

We first show (18) for the sub-optimal arm $i = 2$. Let $x_T = \lfloor \log^{1+\gamma}(T) \rfloor$ with fixed $\gamma \in (0,1)$. Also, let $\delta \in (0, \mu_1 - \mu_2)$. We have the following bounds:

$$\begin{aligned}
\mathbb{P}_{\nu\pi}(N_2(T) > x_T) &\leq \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (\tau_2(x_T), T] \;\text{ s.t. }\; \widetilde{U}_1(N_1(t-1), t-1) \leq \widetilde{U}_2(N_2(t-1), t-1) \Big) \qquad (59) \\
&\leq \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widetilde{U}_1(N_1(t-1), x_T) \leq \widetilde{U}_2(x_T, T) \Big) \\
&\leq \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widetilde{U}_1(N_1(t-1), x_T) \leq \mu_2 + \delta \Big) \qquad (60) \\
&\quad + \mathbb{P}_{\nu\pi}\Big(\widetilde{U}_2(x_T, T) > \mu_2 + \delta \Big). \qquad (61)
\end{aligned}$$

Note that (59) 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.

For the term in (60), we have

$$\begin{aligned}
(60) &\leq \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\left(\widetilde{U}_1(m, x_T) \leq \mu_2 + \delta\right) && (62)\\
&= \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\left(d_P(\widehat{\mu}_1(\tau_1(m)), \mu_2 + \delta) \geq \frac{\log(x_T)}{m},\ \ \widehat{\mu}_1(\tau_1(m)) \leq \mu_2 + \delta\right)\\
&= \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\left(\frac{1}{m}\sum_{l=1}^{m} X_1(l) \leq y_m^*\right)\\
&\leq \sum_{m=1}^{\infty} \exp\left(-m \cdot d_P(y_m^*, \mu_1)\right), && (63)
\end{aligned}$$

where for each m, y_m^* is the unique solution to $d_P(y_m^*, \mu_2 + \delta) = \log(x_T)/m$ and $y_m^* < \mu_2 + \delta$, and we have used a large deviations upper bound in (63). We define

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

and so for $m \geq s_T$,

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

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 $m \geq s_T$,

$$d_P(y_m^*, \mu_1) \geq \frac{\log(x_T)}{m} \cdot \inf_{z < \mu_2 + \delta} \frac{d_P(z, \mu_1)}{d_P(z, \mu_2 + \delta)} + \frac{d_P(\mu_2 + \delta, \mu_1)}{2}. \tag{64}$$

Splitting the sum in (63) into two pieces at $\lfloor s_T \rfloor$, we have

$$\begin{aligned}
(63) &= \sum_{m=1}^{\lfloor s_T \rfloor} \exp\left(-m \cdot d_P(y_m^*, \mu_2 + \delta) \cdot \frac{d_P(y_m^*, \mu_1)}{d_P(y_m^*, \mu_2 + \delta)}\right)\\
&\quad + \sum_{m=\lfloor s_T \rfloor + 1}^{\infty} \exp\left(-m \cdot d_P(y_m^*, \mu_1)\right)\\
&\leq \sum_{m=1}^{\lfloor s_T \rfloor} \exp\left(-m \cdot \frac{\log(x_T)}{m} \cdot \inf_{z < \mu_2 + \delta} \frac{d_P(z, \mu_1)}{d_P(z, \mu_2 + \delta)}\right) && (65)\\
&\quad + \sum_{m=\lfloor s_T \rfloor + 1}^{\infty} \exp\left(-m \cdot \left(\frac{\log(x_T)}{m} \cdot \inf_{z < \mu_2 + \delta} \frac{d_P(z, \mu_1)}{d_P(z, \mu_2 + \delta)} + \frac{d_P(\mu_2 + \delta, \mu_1)}{2}\right)\right) && (66)\\
&= x_T^{-\inf_{z < \mu_2 + \delta} \frac{d_P(z, \mu_1)}{d_P(z, \mu_2 + \delta)}} \cdot \left(\lfloor s_T \rfloor + \sum_{m=\lfloor s_T \rfloor + 1}^{\infty} \exp\left(-m \cdot \frac{d_P(\mu_2 + \delta, \mu_1)}{2}\right)\right). && (67)
\end{aligned}$$

In (65), we use the fact that $d_P(y_m^*, \mu_2 + \delta) = \log(x_T)/m$. In (66), we use (64) (for $m \geq s_T$).

For the term in (61), since $\lim_{T\to\infty} f(T)/x_T = 0$, 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,

$$\begin{aligned}(61) &\le \mathbb{P}_{\nu\pi}\left(\frac{1}{x_T}\sum_{l=1}^{x_T} X_2(l) > \mu_2 + \frac{\delta}{2}\right)\\ &\le \exp\Big(-x_T \cdot \Lambda^*_{P^{\mu_2}}(\mu_2 + \delta/2)\Big),\end{aligned} \tag{68}$$

where $\Lambda^*_{P^{\mu_2}}$ is the large deviations rate function for P^{μ_2}.

Using (60) and (67) together with (61) and (68), we have

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}\left(N_2(T) > x_T\right)}{\log(x_T)} \le -\inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)}. \tag{69}$$

From the argument included separately in Appendix C,

$$\lim_{\delta\downarrow 0} \inf_{z<\mu_2+\delta} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2+\delta)} = \inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2)}. \tag{70}$$

Recall we also have the lower bound:

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}\left(N_2(T) > (1-\gamma)T\right)}{\log((1-\gamma)T)} \ge -\inf_{z<\mu_2} \frac{d_P(z,\mu_1)}{d_P(z,\mu_2)}, \tag{71}$$

as established in the proof of Theorem 1. For the case $i = 2$, the convergence:

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

at the endpoints $x = x_T$ and $x = (1-\gamma)T$ follows from the upper bound in (69)-(70), the lower bound in (71), together with the monotonicity of the function $x \mapsto \log \mathbb{P}_{\nu\pi}\left(N_2(T) > x\right)/\log(x)$ (for any fixed T). The uniform convergence in (72) for $x \in [x_T, (1-\gamma)T]$ follows from the same monotonicity property.

We now show (18) for sub-optimal arm $i \ge 3$. Let $\delta \in (0, \mu_{i-1} - \mu_i)$. In place of (60) and (61), we now have

$$\mathbb{P}_{\nu\pi}\left(N_i(T) > x_T\right) \le \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \ \text{ s.t. } \max_{j\,:\,j\le i-1} \widetilde{U}_j(N_j(t), x_T) \le \mu_i + \delta\Big) \tag{73}$$

$$+ \mathbb{P}_{\nu\pi}\Big(\widetilde{U}_i(x_T, T) > \mu_i + \delta\Big). \tag{74}$$

We can bound (73) via:

$$\begin{aligned}(73) &\le \mathbb{P}_{\nu\pi}\Big(\forall\, j \le i-1, \ \ \exists\, m_j \in \mathbb{Z}_+ \ \text{ s.t. } \ \widetilde{U}_j(m_j, x_T) \le \mu_i + \delta\Big)\\ &\le \prod_{j=1}^{i-1} \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\Big(\widetilde{U}_j(m, x_T) \le \mu_i + \delta\Big),\end{aligned} \tag{75}$$

where (75) follows from the independence of the rewards from different arms. We can then upper bound each term in the product of (75) in the same way as (62). We can upper bound (74) in the same way as (61), and thus show that it is negligible as $T \to \infty$. Following the rest of the argument above (which was for the case $i = 2$), we eventually obtain (due to the product structure in (75)):

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}\left(N_i(T) > x_T\right)}{\log(x_T)} \leq -\sum_{j=1}^{i-1} \inf_{z<\mu_i} \frac{d_P(z,\mu_j)}{d_P(z,\mu_i)}. \tag{76}$$

For sub-optimal arm $i \geq 3$, we also have the lower bound:

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}\left(N_i(T) > (1-\gamma)T\right)}{\log((1-\gamma)T)} \geq -\sum_{j=1}^{i-1} \inf_{z<\mu_i} \frac{d_P(z,\mu_j)}{d_P(z,\mu_i)}, \tag{77}$$

as established in the proof of Theorem 1. For sub-optimal arm $i \geq 3$, the convergence in (72) at the endpoints $x = x_T$ and $x = (1-\gamma)T$ follows from the upper bound in (76), the lower bound in (77), together with the monotonicity of the function $x \mapsto \log \mathbb{P}_{\nu\pi}\left(N_i(T) > x\right) / \log(x)$ (for any fixed T). The uniform convergence in (72) for $x \in [x_T, (1-\gamma)T]$ follows from the same monotonicity property. □

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.

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 between the amount of UCB exploration (in Algorithm 1) and the resulting exponent of the regret distribution tail, as established in (41) and described in Remark 6.

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 9, here we do not have upper bounds on regret tail probabilities

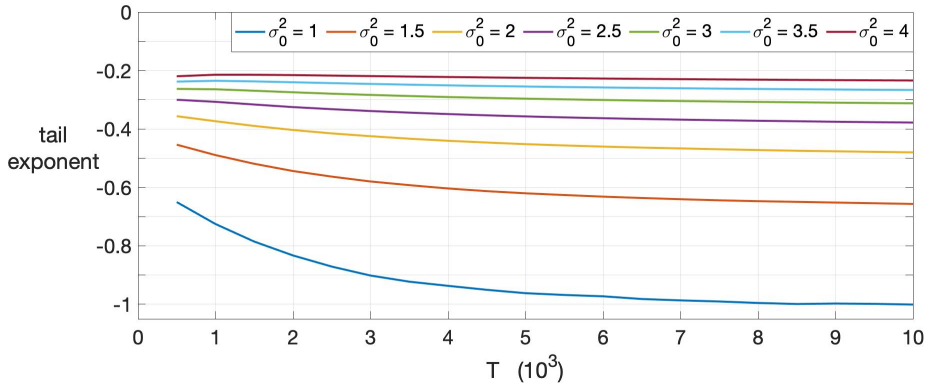

Figure 1 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) \geq 0.8T)/\log(T)$ vs T. Environment $\nu = (N(0.1, \sigma_0^2), N(0, \sigma_0^2))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards. The curves correspond to the cases $\sigma_0^2 = 1, 1.5, \ldots, 4$, as indicated by the legend. The curves asymptote to $-1/\sigma_0^2$ in each case, which agrees with Corollary 2 and (32).

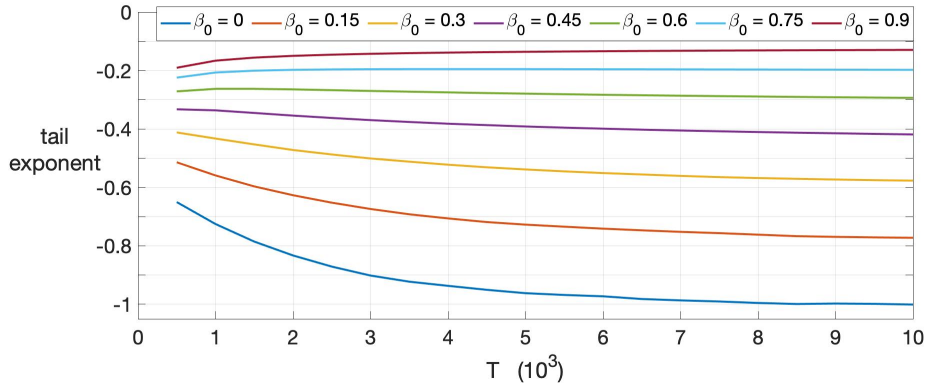

Figure 2 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) \geq 0.8T)/\log(T)$ vs T. Environment ν consists of two Gaussian AR(1) processes with common AR coefficient β_0, and equilibrium distributions $(N(0.1, 1), N(0, 1))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards. The curves correspond 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 (37).

(only lower bounds in (37)), 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 misspecification in this setting. The $\frac{1+\beta_0}{1-\beta_0}$ factor in Figure 5 is taken from the lower bound in (37), which we essentially confirm to be tight here.

Appendix A: Proofs for Section 3.1

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

$$P_\theta(dx) = \exp\left(\theta \cdot x - \Lambda_P(\theta)\right) P(dx), \quad \theta \in \Theta_P. \tag{78}$$

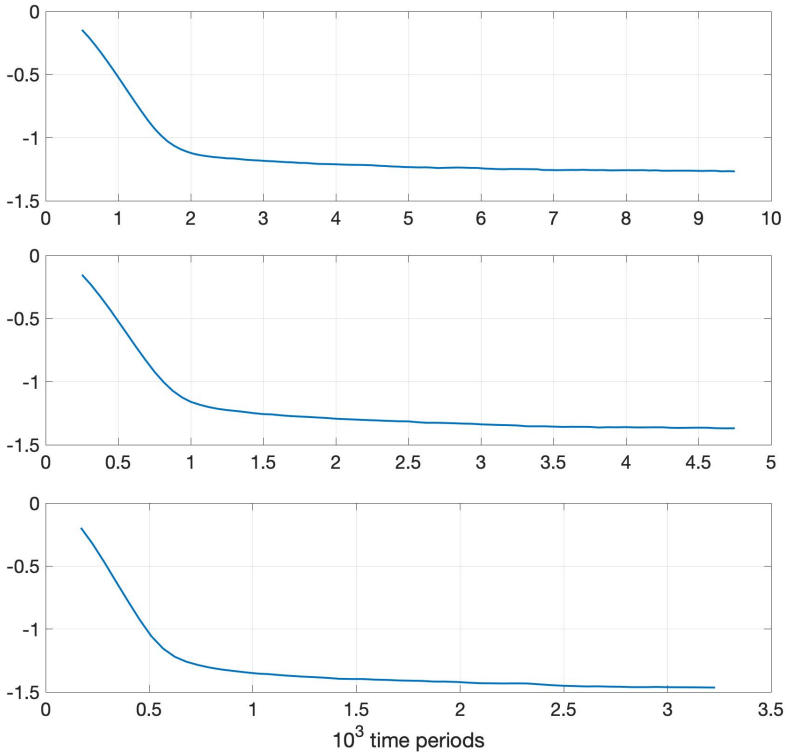

Figure 3 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$ (with time horizon T fixed). Environment $\nu = (\mathrm{Ber}(q), \mathrm{Ber}(0.4))$. Algorithm π is KL-UCB for iid Bernoulli rewards. Top: $q = 0.475$, $T = 10^4$; Middle: $q = 0.5$, $T = 5 \times 10^3$; 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).

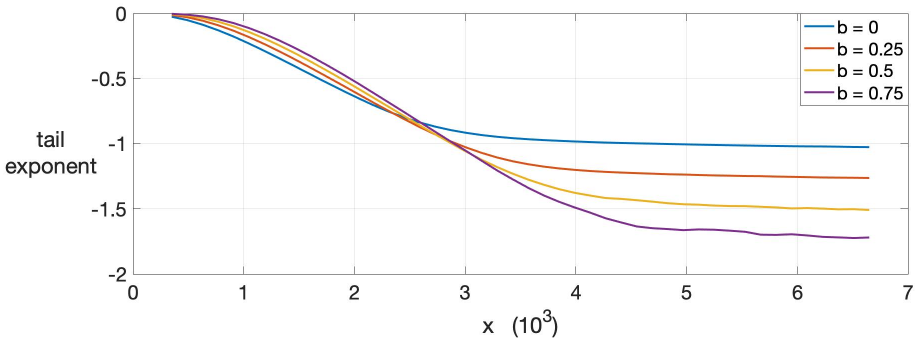

Figure 4 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$, with fixed time horizon $T = 7 \times 10^3$. Environment $\nu = (N(0.1, 1), N(0, 1))$. π is Algorithm 1 with KL divergence d_P between unit-variance Gaussian distributions, and we aim for a regret tail exponent 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 (41), the curves asymptote to -1, -1.25, -1.5, -1.75.

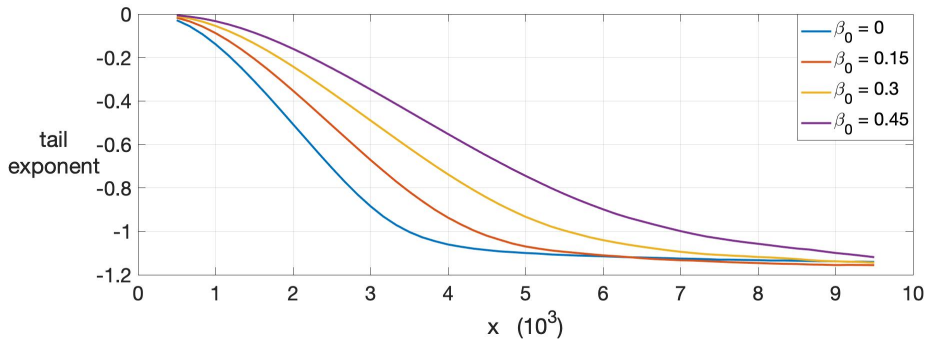

Figure 5 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$, with fixed time horizon $T = 10^4$. Environment ν consists of two Gaussian AR(1) processes with common AR coefficient β_0, and equilibrium distributions $(N(0.1, 1), N(0, 1))$. π is Algorithm 1 with KL divergence d_P between unit-variance Gaussian distributions, and $1 + b = 1.1 \cdot \frac{1+\beta_0}{1-\beta_0}$ (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.

Then the KL divergence between distributions P_{θ_1} and P_{θ_2} has the expression:

$$D(P_{\theta_1} \,\|\, P_{\theta_2}) = \Lambda_P(\theta_2) - \Lambda_P(\theta_1) - \Lambda_P'(\theta_1) \cdot (\theta_2 - \theta_1). \tag{79}$$

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 (78)). For any $\theta_1, \theta_2 \in \Theta_P$ with $\theta_1 > \theta_2$,

$$\inf_{\theta \in \Theta_P : \theta < \theta_2} \frac{D(P_\theta \,\|\, P_{\theta_1})}{D(P_\theta \,\|\, P_{\theta_2})} = 1. \tag{80}$$

We first show the forward direction, that (80) implies (9). Suppose $\inf \Theta_P > -\infty$. Note that (80) implies that for any fixed $\theta_0 > \inf \Theta_P$,

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

Then taking θ_0 arbitrarily close to $\inf \Theta_P$, we must have

$$\lim_{\theta \downarrow \inf \Theta_P} \Lambda_P(\theta) + \epsilon \Lambda_P'(\theta) = -\infty \tag{81}$$

for any $\epsilon > 0$. Because $\inf \Theta_P > -\infty$, (81) implies that

$$\lim_{\theta \downarrow \inf \Theta_P} \Lambda_P'(\theta) = -\infty, \tag{82}$$

since Λ_P is strictly convex and Λ_P' is strictly increasing on Θ_P. Since

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

(81)-(82) imply that

$$\lim_{\theta \downarrow \inf \Theta_P} \left| \frac{\Lambda_P(\theta)}{\Lambda'_P(\theta)} \right| = 0.$$

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 (80) if $\inf \Theta_P > -\infty$. Hence, it must be that $\inf \Theta_P = -\infty$.

Now suppose that

$$\lim_{\theta \to -\infty} \theta \Lambda'_P(\theta) - \Lambda_P(\theta) < \infty. \tag{83}$$

Again, consider two the possible cases:

1. $\lim_{\theta \to -\infty} \Lambda'_P(\theta) = -\infty$
2. $\lim_{\theta \to -\infty} \Lambda'_P(\theta) > -\infty$.

In the first case, (80) cannot hold because (83) implies (for $\theta_1 > \theta_2$):

$$\lim_{\theta \to -\infty} \frac{D(P_\theta \,\|\, P_{\theta_1})}{D(P_\theta \,\|\, P_{\theta_2})} = \lim_{\theta \to -\infty} \frac{\Lambda'_P(\theta)\theta_1}{\Lambda'_P(\theta)\theta_2} = \frac{\theta_1}{\theta_2} \neq 1.$$

In the second case, (80) cannot hold because (83) then implies that $\lim_{\theta \to -\infty} D(P_\theta \,\|\, P_{\theta_0}) < \infty$ for any $\theta_0 \in \Theta_P$. So it must be that

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

Thus, the forward direction is established.

We now show the reverse direction, that (9) implies (80). For any $\theta_1, \theta_2 \in \Theta_P$ fixed with $\theta_1 > \theta_2$,

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

There are 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, note that for any fixed non-zero θ_0, we have

$$\lim_{\theta \to -\infty} \frac{\theta_0 \Lambda'_P(\theta)}{\theta \Lambda'_P(\theta) - \Lambda_P(\theta)} = \lim_{\theta \to -\infty} \frac{\theta_0}{\theta} = 0.$$

And if $\theta_0 = 0$, then of course the same limit result holds. So (9) implies that

$$\lim_{\theta \to -\infty} \frac{D(P_\theta \,\|\, P_{\theta_1})}{D(P_\theta \,\|\, P_{\theta_2})} = \lim_{\theta \to -\infty} \frac{\theta \Lambda'_P(\theta) - \Lambda_P(\theta)}{\theta \Lambda'_P(\theta) - \Lambda_P(\theta)} = 1. \tag{84}$$

In the second case, (9) directly implies (84). 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 using $\lim_{\theta \to -\infty} \Lambda'_P(\theta) = -\infty$ yields the desired result. $\square$

Proof of Proposition 2. Let X be a random variable with distribution P. We first address the case where the distributions assign no mass to the (finite) infimum of their support, which we denote by L. For some $l > L$ with $l - L$ small (which will be made precise later), we have by the definition of convex conjugation:

$$\begin{aligned} \Lambda^*_P(l) &= \sup_{\theta \in \Theta_P} \left(\theta \cdot l - \log \mathbb{E}[\exp(\theta X)] \right) \\ &= -\log \left(\inf_{\theta \in \Theta_P} \mathbb{E}[\exp(\theta(X - l))] \right). \end{aligned} \tag{85}$$

Let us decompose via:

$$\mathbb{E}[\exp(\theta(X-l))] = \mathbb{E}[\exp(\theta(X-l)); X \geq l] + \mathbb{E}[\exp(\theta(X-l)); X < l]. \tag{86}$$

For $\theta < 0$, the first term on the right side of (86) can be bounded via:

$$0 \leq \mathbb{E}[\exp(\theta(X-l)); X \geq l] \leq \exp(|\theta(l-L)|) \cdot \mathbb{E}[\exp(\theta(X-L))], \tag{87}$$

while the second term can be bounded via:

$$0 \leq \mathbb{E}[\exp(\theta(X-l)); X < l] \leq \exp(|\theta(l-L)|) \cdot \mathbb{P}(X < l). \tag{88}$$

Now for any $\epsilon > 0$, set $\theta = -1/\epsilon$ and $l - L = \epsilon$, so that $\exp(|\theta(l-L)|) = \exp(1)$ in (87)-(88). Since $X \geq L$ with probability one and X always has continuous CDF in a neighborhood of L, by the bounded convergence theorem, $\lim_{\theta \to -\infty} \mathbb{E}[\exp(\theta(X-L))] = 0$. Moreover, $\lim_{l \downarrow L} \mathbb{P}(X < l) = 0$. So

the upper bounds in (87)-(88) can be made arbitrarily small by taking (the just defined) $\epsilon > 0$ to be sufficiently small. Therefore, we have shown that

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

which, by (85), translates into

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

Since

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

we have

$$\lim_{\theta \to -\infty} \Lambda_P^*(\Lambda_P'(\theta)) = \infty, \tag{89}$$

which is the equivalent representation for (9).

In the case where there is strictly positive mass on L, note that if we take $l > L$ with l sufficiently close to L, then the infimum in $\inf_{\theta \in \Theta_P} \mathbb{E}[\exp(\theta(X-l))]$ from (85) is achieved with $\theta < 0$. (This is completely trivial if $\mathbb{P}(X = L) = \mathbb{P}(X < l)$ for $l > L$ with l sufficiently close to L.) So it suffices to simply consider $\theta < 0$. Then for $l > L$ with l sufficiently close to L, we have the lower bounds:

$$\begin{aligned} \inf_{\theta \in \Theta_P} \mathbb{E}[\exp(\theta(X-l))] &\geq \inf_{\theta < 0} \mathbb{E}[\exp(\theta(X-l)); X = L] \\ &= \inf_{\theta < 0} \exp(\theta(L-l)) \cdot \mathbb{P}(X = L) \\ &= \mathbb{P}(X = L). \end{aligned}$$

Therefore, we have shown that

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

which, by (85), translates into

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

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

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

unlike in the case of continuous distributions, where we ended up with (89), here we have

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

□

Appendix B: Proofs for Section 3.2

Proof of Proposition 4. Let i be any sub-optimal arm. From the lower bounds in (55) 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,

$$\begin{aligned} 0 &\le \mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_i(T) - \mu_i| > \epsilon \mid N_i(T) > x\right) \\ &\le \frac{\mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_i(T) - \mu_i| > \epsilon,\, N_i(T) > \log^{1+\gamma}(T)\right)}{\mathbb{P}_{\nu\pi}\left(N_i(T) > (1-\gamma)T\right)} \\ &\le T^a \cdot 2\exp\left(-\log^{1+\gamma}(T) \cdot (\Lambda^*_{P^{\mu_i}}(\mu_i + \epsilon) \wedge \Lambda^*_{P^{\mu_i}}(\mu_i - \epsilon))\right), \end{aligned}$$

where to obtain the last inequality, we use Cramér's Theorem (see, for example, Theorem 2.2.3 on page 27 of Dembo and Zeitouni (1998)) to upper bound the numerator. 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. □

Appendix C: Proofs for Section 3.3

Verification of (70) in Proof of Theorem 2. With the natural parameterization of an exponential family P_θ, $\theta \in \Theta_P$, as in (78), with KL divergence as in (79), 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 D: Proofs for Section 3.4

The proof of Lemma 2 is a simplification of the proof of Proposition 5.

LEMMA 2. *Under the assumptions of Theorem 3, for any environment $\nu = (P_1,\dots,P_K) \in \mathcal{M}^K$ 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 2. Suppose the conclusion is false for some environment $\widetilde{\nu} = (\widetilde{P}_1, P_2, \dots, P_K) \in \mathcal{M}^K$. 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

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(N_1(T_n) \le m\right) \ge \epsilon. \tag{90}$$

Denote the event in (90) by $\mathcal{A}'_n$. Consider another environment $\nu = (P_1, P_2, \dots, P_K) \in \mathcal{M}^K$ where arm 1 is optimal (with all other arms being the same as in $\widetilde{\nu}$). Pick $L > 0$ large enough so that

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(\forall\, l = 1,\dots,m : \frac{1}{l}\sum_{t=1}^{l} \log \frac{dP_1}{d\widetilde{P}_1}(X_1(t)) \ge -D(\widetilde{P}_1 \,\|\, P_1) - L\right) \ge 1 - \epsilon/2. \tag{91}$$

Define

$$\mathcal{B}'_n = \left\{ \frac{1}{N_1(T_n)} \sum_{t=1}^{N_1(T_n)} \log \frac{dP_1}{d\widetilde{P}_1}(X_1(t)) \ge -D(\widetilde{P}_1 \,\|\, P_1) - L \right\}.$$

Following the same steps from (23)-(25) but with $\mathcal{A}'_n$, $\mathcal{B}'_n$ in the place of $\mathcal{A}_n$, $\mathcal{B}_n$, respectively,

$$\mathbb{P}_{\nu\pi}(\exists\, i \neq 1 : N_i(T_n) > T_n/(2K)) \ge \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}'_n, \mathcal{B}'_n) \cdot \exp\left(-\left(D(\widetilde{P}_1 \,\|\, P_1) + L\right) m\right).$$

From (90) and (91), we have $\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}'_n, \mathcal{B}'_n) \ge \epsilon/2$ for all n. This violates the $\mathcal{M}$-consistency of π, and thus (90) cannot be true. □

Appendix E: Proofs for Section 4.1

Proof of Proposition 7. The proof of the lower bound part of (31) follows from Theorem 4 (which uses Proposition 8). To establish the upper bound part of (31), we can use the same proof of Theorem 2; see Section 6.2. Without loss of generality, suppose that $\mu(Q_1) > \mu(Q_2) > \cdots > \mu(Q_K)$ (i.e., $r(i) = i$ for all $i \in [K]$). So the only thing that needs to be checked is the analog of (70):

$$\lim_{\delta \downarrow 0} \inf_{z < \mu(Q_2)+\delta} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2)+\delta)} = \inf_{z < \mu(Q_2)} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2))}. \tag{92}$$

(Below, we check (92) 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:

$$\inf_{z < \mu(Q_2)} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2))} = \inf_{z < \mu(Q_2)-\eta} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2))}, \tag{93}$$

$$\inf_{z < \mu(Q_2)+\delta} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2)+\delta)} = \inf_{z < \mu(Q_2)-\eta} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2)+\delta)}. \tag{94}$$

Note that

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

is monotone decreasing for $z < \mu(P_2)$, which we deduce from the verification of (70) in proof of Theorem 2 in Appendix C. Also, we have:

$$\lim_{\delta \downarrow 0} \sup_{z < \mu(Q_2)-\eta} \frac{d_P(z, \mu(Q_2))}{d_P(z, \mu(P_2)+\delta)} = \sup_{\delta > 0} \sup_{z < \mu(Q_2)-\eta} \frac{d_P(z, \mu(Q_2))}{d_P(z, \mu(Q_2)+\delta)} = 1,$$
$$\lim_{\delta \downarrow 0} \frac{d_P(\mu(Q_2)-\eta, \mu(Q_2))}{d_P(\mu(Q_2)-\eta, \mu(Q_2)+\delta)} = 1.$$

Therefore, we have uniform convergence for $z < \mu(P_2) - \eta$:

$$\lim_{\delta \downarrow 0} \sup_{z < \mu(Q_2)-\eta} \left| \frac{d_P(z, \mu(Q_2))}{d_P(z, \mu(Q_2)+\delta)} - 1 \right| = 0. \tag{95}$$

For any $\epsilon \in (0,1)$, using (95), we have for sufficiently small $\delta > 0$:

$$\begin{aligned}(1-\epsilon) \inf_{z < \mu(Q_2)-\eta} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2))} &\leq \inf_{z < \mu(Q_2)-\eta} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2)+\delta)} \cdot \frac{d_P(z, \mu(Q_2))}{d_P(z, \mu(Q_2))} \\ &\leq (1+\epsilon) \inf_{z < \mu(Q_2)-\eta} \frac{d_{Q_1}(z, \mu(Q_1))}{d_P(z, \mu(Q_2))}.\end{aligned}$$

Sending $\delta \downarrow 0$, followed by $\epsilon \downarrow 0$, and then using (93)-(94), we obtain (92). $\square$

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 $i = 1, 2$). The proof of the lower bound part of (31) follows from Theorem 4 (which uses Proposition 8). The upper bound part:

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > \log^{1+\gamma}(T))}{\log(\log^{1+\gamma}(T))} \le -\frac{\sigma^2}{\sigma_0^2}, \tag{96}$$

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 uniformity over x follows by the monotonicity of $x \mapsto \log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ for fixed T and $x > 1$.

□

Verification of (96) 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:

$$\begin{aligned}
&\mathbb{P}_{\nu\pi}\left(N_2(T) > x_T\right) \\
&\le \mathbb{P}_{\nu\pi}\left(\exists\, t \in (\tau_2(x_T), T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(t-1)}{N_1(t-1)}} \le \widehat{\mu}_2(t-1) + \sqrt{\frac{2\sigma^2 \log(t-1)}{N_2(t-1)}}\right) \\
&\le \mathbb{P}_{\nu\pi}\left(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(x_T)}{N_1(t-1)}} \le \widehat{\mu}_2(\tau_2(x_T)) + \sqrt{\frac{2\sigma^2 \log(T)}{x_T}}\right) \\
&\le \mathbb{P}_{\nu\pi}\left(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(x_T)}{N_1(t-1)}} \le \mu_2 + \frac{\Delta}{2}\right) \qquad (97) \\
&\quad + \mathbb{P}_{\nu\pi}\left(\widehat{\mu}_2(\tau_2(x_T)) + \sqrt{\frac{2\sigma^2 \log(T)}{x_T}} > \mu_2 + \frac{\Delta}{2}\right). \qquad (98)
\end{aligned}$$

For the term in (97), we have

$$\begin{aligned}
(97) &= \mathbb{P}_{\nu\pi}\left(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(x_T)}{N_1(t-1)}} \le \mu_1 - \frac{\Delta}{2}\right) \\
&\le \sum_{m=1}^{\infty} \mathbb{P}_{\nu\pi}\left(\frac{1}{m}\sum_{l=1}^{m} X_1(l) \le \mu_1 - \sqrt{\frac{2\sigma^2 \log(x_T)}{m}} - \frac{\Delta}{2}\right) \qquad (99) \\
&\le \sum_{m=1}^{\infty} \exp\left(-\frac{m}{2\sigma_0^2}\left(\sqrt{\frac{2\sigma^2 \log(x_T)}{m}} + \frac{\Delta}{2}\right)^2\right) \qquad (100) \\
&= x_T^{-\sigma^2/\sigma_0^2} \cdot \sum_{m=1}^{\infty} \exp\left(-\frac{\sqrt{m\sigma^2 \log(x_T)}\Delta}{\sqrt{2}\sigma_0^2} - \frac{m\Delta^2}{8\sigma_0^2}\right) \\
&\le x_T^{-\sigma^2/\sigma_0^2} \cdot \sum_{m=1}^{\infty} \exp\left(-\frac{\sqrt{m}\sigma\Delta}{\sqrt{2}\sigma_0^2} - \frac{m\Delta^2}{8\sigma_0^2}\right) \qquad (\text{for } T \ge 16), \qquad (101)
\end{aligned}$$

where to obtain (99), we have used a union bound over all possible values of $N_1(t)$, $t \geq 1$, and to obtain (100), we have used a large deviations upper bound.

For the term in (98), we have for sufficiently large T,

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

So for sufficiently large T,

$$\begin{aligned}(98) &\leq \mathbb{P}_{\nu\pi}\left(\frac{1}{x_T}\sum_{t=1}^{x_T} X_2(t) > \mu_2 + \frac{\Delta}{4}\right) \\ &\leq \exp\left(-x_T \cdot \frac{\Delta^2}{32\sigma_0^2}\right), \end{aligned} \tag{102}$$

where to obtain (102), we have used a large deviations upper bound.

Putting together (97), (101) and (98), (102), 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 F: Proofs for Section 4.2

Proof of Proposition 8. 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

$$\left\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} X_i(t) = \mu_i,\ i \in [K]\right\}. \tag{103}$$

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

$$U_i(t) = \sup\left\{z \in \mathcal{I}_P : d_P(\widehat{\mu}_i(t), z) \leq \frac{\log(t)}{N_i(t)}\right\}, \tag{104}$$

where, as defined previously, $\widehat{\mu}_i(t) = \frac{1}{N_i(t)}\sum_{s=1}^{N_i(t)} X_i(s)$.

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

$$N_i(T) = 1 + \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ U_i(t) \geq \mu_1 - \delta,\ \widehat{\mu}_i(t) \leq \mu_i + \delta\right) \tag{105}$$

$$+ \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ U_i(t) \geq \mu_1 - \delta,\ \widehat{\mu}_i(t) > \mu_i + \delta\right) \tag{106}$$

$$+ \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = i,\ U_i(t) < \mu_1 - \delta\right), \tag{107}$$

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

The first sum is upper bounded via:

$$(105) \leq \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1)=i,\, d_P(\mu_i+\delta, \mu_1-\delta) \leq \frac{\log(t)}{N_i(t)}\right) \tag{108}$$

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

$$\leq \frac{\log(T)}{d_P(\mu_i+\delta, \mu_1-\delta)} + 1. \tag{109}$$

The bound in (108) 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 (104).

The second sum is upper bounded via:

$$(106) \leq \sum_{t=K}^{\infty} \mathbb{I}\left(A(t+1)=i,\, \widehat{\mu}_i(t) > \mu_i + \delta\right). \tag{110}$$

On sample paths in (103), the indicators on the right side of (110) can equal 1 for at most finitely many t. (For each 1 in the sum, arm i is played an additional time and an additional sample incorporated into $\widehat{\mu}_i(t)$.)

The third sum is upper bounded via:

$$(107) \leq \sum_{t=K}^{\infty} \mathbb{I}\left(A(t+1)=i,\, U_1(t) \leq U_i(t) < \mu_1 - \delta\right)$$

$$\leq \sum_{t=K}^{\infty} \mathbb{I}\left(U_1(t) < \mu_1 - \delta\right). \tag{111}$$

On sample paths in (103), the indicators on the right side of (111) can equal 1 for at most finitely many t. (As $t \to \infty$, either $N_1(t)$ increases to infinity or remains finite. 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, $\log(t)$ in (104) 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 (109)-(111), and sending $T \to \infty$ followed by $\delta \downarrow 0$, we have for each sub-optimal arm $i \geq 2$,

$$\limsup_{T\to\infty} \frac{N_i(T)}{\log(T)} \leq \frac{1}{d_P(\mu_i, \mu_1)}. \tag{112}$$

Therefore, for the optimal arm 1,

$$\lim_{T\to\infty} \frac{N_1(T)}{T} = 1, \tag{113}$$

which, by the form of the index in (104), then implies:

$$\lim_{t\to\infty} U_1(t) = \mu_1. \tag{114}$$

Then, (113) and (114) imply that for each sub-optimal arm $i \geq 2$,

$$\lim_{T\to\infty} N_i(T) = \infty. \tag{115}$$

(If (115) is not true for some sub-optimal arm j, then since the term $\log(t)$ grows without bound in the index (104), we would eventually have $U_j(t) > \mu_1 + \epsilon > U_1(t)$ for some $\epsilon > 0$ and all t sufficiently large, thereby contradicting (113).)

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 $i \geq 2$,

$$U_1(\tau_1(m)-1) > U_i(\tau_1(m)-1). \tag{116}$$

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

$$\begin{aligned}
\max_{t\in[\tau_1(m),\tau_1(m+1)]} \frac{\log(t)}{N_i(t)} &\leq \frac{\log(\tau_1(m+1))}{N_i(\tau_1(m)-1)} \\
&= \frac{\log(\tau_1(m+1))}{\log(\tau_1(m)-1)} \frac{\log(\tau_1(m)-1)}{N_i(\tau_1(m)-1)} \\
&\leq (1+\delta)\frac{\log(\tau_1(m)-1)}{N_i(\tau_1(m)-1)} && (117)\\
&\leq (1+\delta) d_P(\mu_i - \delta, U_i(\tau_1(m)-1)) && (118)\\
&\leq (1+\delta) d_P(\mu_i - \delta, U_1(\tau_1(m)-1)) && (119)\\
&\leq (1+\delta) d_P(\mu_i - \delta, \mu_1 + \delta). && (120)
\end{aligned}$$

Note that (117) is due to (113), (118) 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 (104), (119) is due to (116), and (120) is due to (114). From (120), we see that

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

which together with (112), completes the proof. □

Proof of Theorem 4. Without loss of generality, suppose that the long-run average rewards (in the sense of (33)) 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 arm $j \leq i-1$ 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 (35) is empty, and the lower bound is $-\infty$.) So in the environment $\widetilde{\nu}$, arm i yields the highest 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 (51)-(53) in the proof of Theorem 1,

$$\begin{aligned}
&\mathbb{P}_{\nu\pi}(N_i(T) > (1-\gamma)T) \\
&= \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(N_i(T) > (1-\gamma)T\right) \exp\left(\sum_{j=1}^{i-1} \left(-\theta_j \cdot \sum_{t=1}^{N_j(T)} X_j(t) + N_j(T) \cdot \overline{\Lambda}_j^{N_j(T)}(\theta_j) \right) \right) \right] && (121) \\
&\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(\sum_{j=1}^{i-1} \Big(-\theta_j \cdot \widehat{\mu}_j(T) + \overline{\Lambda}_j^{N_j(T)}(\theta_j) \Big) N_j(T) \right) \right] && (122) \\
&\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(\sum_{j=1}^{i-1} \Big(-\theta_j \cdot \big(\overline{\Lambda}_j'(\theta_j) - \delta\big) + \overline{\Lambda}_j(\theta_j) - \delta \Big) N_j(T) \right) \right] && (123) \\
&= \mathbb{E}_{\widetilde{\nu}\pi}\left[\mathbb{I}\left(\mathcal{A}_T, \mathcal{B}_T\right) \exp\left(-\sum_{j=1}^{i-1} \Big(\overline{\Lambda}_j^*(\overline{\Lambda}_j'(\theta_j)) + \delta(1-\theta_j) \Big) N_j(T) \right) \right] && (124) \\
&\geq \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{A}_T, \mathcal{B}_T) \cdot \exp\left(-\sum_{j=1}^{i-1} \Big(\overline{\Lambda}_j^*(\overline{\Lambda}_j'(\theta_j)) + \delta(1-\theta_j) \Big) \left(\frac{1}{d_P(\overline{\Lambda}_j'(\theta_j), \overline{\Lambda}_i'(0))} + \delta \right) \log(T) \right). && (125)
\end{aligned}$$

In (121), we have performed a change-of-measure from environment ν to $\widetilde{\nu}$. In (122), 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 (123), and the relevant identity for the convex conjugates $\overline{\Lambda}_j^*$ in (124). We have used the event $\mathcal{A}_T$ in (125). 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 (34). 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 (125), 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 (35) holds with the infimum over $B_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$ due to $x \mapsto \log \mathbb{P}_{\nu\pi}(N_i(T) > x)/\log(x)$ being monotone decreasing for $x \in B_\gamma(T)$, with T fixed. □

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* .

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).

Fan L, Glynn P (2022) The typical behavior of bandit algorithms. *Working Paper* .

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.

Garivier A, Menard P, Stoltz G (2019) Explore first, exploit next: the true shape of regret in bandit problems. *Mathematics of Operations Research* 44(2):377–399.

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

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.

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).

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* .

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* .

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

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* .