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

---

Contents

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). One main approach to algorithm design is to build 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*. Two important examples of such optimized algorithms are the KL-upper confidence bound (KL-UCB) algorithm and Thompson sampling (TS); see Cappé et al. (2013) and Korda et al. (2013).

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 if $0 < c < 1$, then $\mathbb{P}(R(T) \geq cT) = O(\log(T)/T)$ as $T \to \infty$. One of our central results is a lower bound characterization of $\mathbb{P}(R(T) \geq cT)$ 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

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

as $T \to \infty$, uniformly in x with $T^a \leq x \leq cT$ for any $0 < a < 1$ and $0 < c < 1$. (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 quite 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 4).

A final sense in which such optimized algorithms are fragile is that when one only slightly modifies the objective, the behavior of the optimized 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 algorithm, 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 a desired $b > 0$, then the tail of the resulting regret distribution will have an exponent of $-(1+b)$ (or less).

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, Assumption 2, which holds essentially for all continuous reward distributions. We provide an intuitive proof sketch for Theorem 1 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 not truncated Cauchy (because Assumption 2 does not hold). In Section 3.4, we provide an alternative, intuitive proof of the generalized Lai-Robbins lower bound for expected regret by focusing on the regret tail and using our change-of-measure arguments. Our new proof differs from existing proofs in the literature and sheds new light on the result. 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 results make use of general lower bounds for the regret tail of KL-UCB when the rewards come from stochastic processes, which we establish in Section 4.2. 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. We provide the formal proofs of Theorems 1 and 2 in Sections 6.1 and 6.2, respectively. We conclude with numerical experiments in Section 7.

1.1. Related Work

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

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

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

and the CLT:

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

where $\Delta_k > 0$ is the difference between the mean of the optimal arm k^* and that of sub-optimal arm k, 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 is a collection of distributions $\nu = (P_1, \ldots, P_K)$, which we refer to as a bandit environment, where each P_k 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 $P_{A(t)}(\cdot)$. We write $X_k(t)$ to denote the reward received when arm k is played for the t-th instance, so that $Y(t) = X_{A(t)}(N_{A(t)}(t))$, where $N_k(t) = \sum_{i=1}^{t} \mathbb{I}(A(i) = k)$ denotes the number of plays of arm k up to and including time t.

We assume that there exists a dominating measure η on $\mathbb{R}$ such that for each arm k, P_k is absolutely continuous with respect to η with density p_k. For any time n, the interaction between the algorithm π and the environment ν induces a probability distribution on $([K] \times \mathbb{R})^n$ with density (with respect to the product measure $(\chi \times \eta)^n$, where χ is counting measure):

$$p^{\nu\pi}(a_1, y_1, \ldots, a_n, y_n) = \prod_{t=1}^{n} \pi_t(a_t \mid a_1, y_1, \ldots, a_{t-1}, y_{t-1}) p_{a_t}(y_t).$$

We write $\mathbb{E}_{\nu\pi}[\cdot]$ to denote the expectation and $\mathbb{P}_{\nu\pi}(\cdot)$ to denote the probability under the distribution corresponding to density $p^{\nu\pi}$.

To design an algorithm for a K-armed bandit, we first specify a model $\mathcal{M}$ of possible arm reward distributions. This induces a class $\mathcal{M}^K$ of bandit environments, with each environment consisting of a K-tuple of arm reward distributions from $\mathcal{M}$. The algorithm is then designed for environments from the class $\mathcal{M}^K$. In the MAB literature, it is almost always assumed that the true environment ν belongs to the specified class $\mathcal{M}^K$. We will take this view in Section 3. In Section 4, we will consider settings where ν lies outside of $\mathcal{M}^K$.

The quality of an algorithm π operating in an environment $\nu = (P_1, \ldots, P_K)$ is measured by the (pseudo-)regret (at time T):

$$R(T) = \sum_k N_k(T) \Delta_k,$$

where $\Delta_k = \mu_* - \mu(P_k)$ and $\mu_* = \max_k \mu(P_k)$. (For a distribution P, we use $\mu(P)$ to denotes its mean.) An arm k is called optimal if $\Delta_k = 0$, and sub-optimal if $\Delta_k > 0$. 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 need to explicitly reference the i-th best arm for $1 \leq i \leq K$. We will use $k(i) \in [K]$ to denote the index (or label) of the i-th best arm. When using this ordering convention, we will assume that the i-th best arm is unique, i.e., $\mu(P_{k(1)}) > \mu(P_{k(2)}) > \cdots > \mu(P_{k(K)})$. When not using this ordering convention, unless specified otherwise, the best arm(s), the second-best arm(s), the third-best arm(s), etc., do not need to be unique.

2.2. Optimized Algorithms

In order to discuss optimized algorithms, we consider arm reward distributions from an exponential family with densities of the form:

$$p(x,\theta) = h(x)\exp\left(\theta \cdot x - A(\theta)\right), \quad \theta \in \Theta, \tag{1}$$
$$\Theta = \left\{\theta \in \mathbb{R} : \int h(x)\exp(\theta \cdot x)\eta(dx) < \infty\right\},$$

with respect to a dominating measure η on $\mathbb{R}$. This parameterization coincides with natural one-dimensional exponential families (see Lehmann and Casella (1998)). (All of our theory easily extends to exponential family models with parameterization: $p(x,\theta) = h(x)\exp\left(\theta \cdot S(x) - A(\theta)\right)$, for monotone sufficient statistic $S(x)$, but we use the model (1) for simplicity.) Without loss of generality, we take h to be a probability density with respect to η, so that (1) corresponds to different exponential tilts of h, with A being the cumulant generating function (CGF) of h. (Note that we can always take $h(x) = p(x,\theta_0)$ for some fixed $\theta_0 \in \Theta$, and then re-define the family of distributions by shifting the parameter values via $\theta \to \theta - \theta_0$ and suitably modifying Θ and A.) Recall that if X is a random variable with density $p(x,\theta)$, then the mean is $\mu(\theta) = \mathbb{E}_\theta[X] = A'(\theta)$, and $\mu(\theta_1) > \mu(\theta_2)$ for $\theta_1 > \theta_2$. Also recall that the KL divergence between distributions from (1) with parameters θ_1 and θ_2 can be expressed as:

$$D(\theta_1,\theta_2) = \mathbb{E}^{\theta_1}\left[\log\frac{p(X,\theta_1)}{p(X,\theta_2)}\right] = A(\theta_2) - A(\theta_1) - A'(\theta_1)\cdot(\theta_2 - \theta_1). \tag{2}$$

In the current context, we assume that the model $\mathcal{M}$ of arm reward distributions consists of an exponential family as in (1). Then, each environment within the class $\mathcal{M}^K$ consists of a K-tuple of exponential family distributions denoted as $(\theta_1,\ldots,\theta_K)$, where arm i has parameter value θ_i. 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 $\mathcal{M}^K$, which is stated as follows. Let π be a so-called *consistent* algorithm, satisfying for any $a > 0$, any environment $\nu \in \mathcal{M}^K$ and any sub-optimal arm k:

$$\lim_{T\to\infty}\frac{\mathbb{E}_{\nu\pi}[N_k(T)]}{T^a} = 0. \tag{3}$$

(The notion of consistency rules out unnatural algorithms which over-specialize and perform very well in particular environments, but very poorly in others.) Then for any environment $\nu = (\theta_1,\ldots,\theta_K) \in \mathcal{M}^K$ and any sub-optimal arm k,

$$\liminf_{T\to\infty}\frac{\mathbb{E}_{\nu\pi}[N_k(T)]}{\log(T)} \geq \frac{1}{D(\theta_k,\theta_*)}, \tag{4}$$

where θ_* denotes the parameter value corresponding to the reward distribution with the maximum mean μ_*. We say that an algorithm π is optimized for $\mathcal{M}^K$ if the lower bound in (4) is achieved, i.e., the condition in Assumption 1 holds.

ASSUMPTION 1. *For any environment* $\nu = (\theta_1, \ldots, \theta_K) \in \mathcal{M}^K$ *and any sub-optimal arm* k,

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_k(T)]}{\log(T)} = \frac{1}{D(\theta_k, \theta_*)}.$$

3. Characterization of the Regret Distribution Tail

3.1. Truncated Cauchy Tails

In this section, we show that for many classes $\mathcal{M}^K$ of exponential family bandit environments (with $\mathcal{M}$ as in (1)), the tail of the regret distribution of optimized algorithms (satisfying Assumption 1) is essentially that of a truncated Cauchy distribution. Moreover, for such classes $\mathcal{M}^K$, 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 $\mathcal{M}$ is a continuous exponential family with left tails that are lighter than exponential (possessing CGF's that are finite on the negative half of the real line). When $\mathcal{M}$ is a discrete exponential family or has exponential left tails, the regret distribution tail is lighter than truncated Cauchy, but it is still heavy and decays 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 Assumption 2 below regarding the exponential family model in (1). It is straightforward to verify this assumption from (1). Following the statement of the theorem, we will provide an equivalent characterization as well as simple sufficient conditions for this assumption.

ASSUMPTION 2. *For the exponential family in (1),* $\inf\Theta = -\infty$ *and* $\lim_{\theta\to-\infty} \theta A'(\theta) - A(\theta) = \infty$. *(Equivalently,* $\lim_{\theta\to-\infty} A^*(A'(\theta)) = \infty$, *where* A^* *is the convex conjugate of* A.*)*

The full proof of Theorem 1 is given in Section 6.1. In Section 3.2, we give a proof sketch for a basic 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 $\mathcal{M}$ be an exponential family as in (1). Suppose the algorithm π is optimized for $\mathcal{M}^K$, i.e., Assumption 1 holds. Then for any environment $\nu = (\theta_1, \ldots, \theta_K) \in \mathcal{M}^K$ and any $i \geq 2$,*

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T) \geq x)}{\log(x)} \geq -\sum_{j=1}^{i-1} \inf_{\theta\in\Theta\,:\,\theta<\theta_{k(i)}} \frac{D(\theta,\theta_{k(j)})}{D(\theta,\theta_{k(i)})} \tag{5}$$

uniformly for $x \in [T^\gamma, (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$.
If in addition Assumption 2 holds, then

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(2)}(T) \geq x)}{\log(x)} = -1 \tag{6}$$

uniformly for $x \in [T^\gamma, (1-\gamma)T]$ for any $\gamma \in (0,1)$ as $T \to \infty$. For any $i \geq 3$, the right side of (5) is equal to $-(i-1)$.

In Lemma 1 below, we provide an equivalent characterization of Assumption 2. This characterization implies that each summand on the right side of (5) is equal to -1. (Note that for any $\theta < \theta' < \theta''$, we have $D(\theta,\theta'')/D(\theta,\theta') \geq 1$.) In light of the exact -1 tail exponent for the second-best arm $k(2)$ in (6), we might conjecture that the lower bounds in (5) are tight in general without Assumption 2. 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. *For the exponential family in (1), Assumption 2 holds if and only if for any fixed $\theta_1 > \theta_2$,*

$$\inf_{\theta\in\Theta\,:\,\theta<\theta_2<\theta_1} \frac{D(\theta,\theta_1)}{D(\theta,\theta_2)} = 1. \tag{7}$$

In Proposition 1, we give a simple sufficient condition for Assumption 2 that applies to reward distributions with support that is unbounded to the left on the real line. The requirement is that the tilting parameter θ in the model (1) can be made arbitrarily negative. In Proposition 2, we provide simple conditions to determine whether or not Assumption 2 holds for distributions with support that is bounded to the left on the real line. When the support is bounded to the left, Assumption 2 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. *For the exponential family in (1), if the support of the distributions is unbounded to the left and $\inf\Theta = -\infty$, then Assumption 2 holds.*

PROPOSITION 2. *For the exponential family in (1), if the support of the distributions is bounded to the left and the distributions assign no mass to the infimum of their support, then Assumption 2 holds. But if the distributions assign strictly positive mass to the infimum of their support, then Assumption 2 does not hold.*

It can be verified that for fixed $\theta_1 > \theta_2$,

$$\inf_{\theta \in \Theta \,:\, \theta < \theta_2 < \theta_1} \frac{D(\theta, \theta_1)}{D(\theta, \theta_2)} = \lim_{\theta \downarrow \inf \Theta} \frac{D(\theta, \theta_1)}{D(\theta, \theta_2)}. \tag{8}$$

The KL divergence $D(\theta, \theta')$ can be thought of as the mean information for discriminating between θ and θ', given a sample from the θ distribution. Since $D(\theta, \theta_1) = D(\theta, \theta_2)$ if and only if $\theta_1 = \theta_2$, the ratio $D(\theta, \theta_1)/D(\theta, \theta_2)$ can be thought of as a measure of the difficulty of discriminating between θ and θ_1 relative to that between θ and θ_2, given a sample from θ 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 (5) 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. Consider a Gaussian family with unknown mean and known (common) variance. Sending $\theta \to -\infty$ induces a location shift of the Gaussian distribution to $-\infty$. Different θ_1 and θ_2 correspond to the same Gaussian distributions, just shifted by some fixed amount. So as we focus on regions increasingly far out in the left tail, it becomes impossible to distinguish between the shifted distributions corresponding to θ_1 and θ_2. From a technical standpoint, $D(\theta, \theta_1) \to \infty$ and $D(\theta, \theta_2) \to \infty$ at the same rate as $\theta \to -\infty$. Hence, (8) is always equal to 1 in this setting.

Example 2. Consider a family of exponential distributions on the negative half of the real line. For example, let $p(x, \theta) = (1 + \theta) e^{(1+\theta)x} \cdot \mathbb{I}(x \leq 0)$ in the model (1), so that $\Theta = (-1, \infty)$. Sending $\theta \downarrow -1$ allocates increasingly more mass increasingly far out in the left tail. However, no matter how far out we move in the left tail, different θ_1 and θ_2 correspond to distributions with different scale parameters, and it will still be possible to distinguish between them. From a technical standpoint, $D(\theta, \theta_1) \to \infty$ and $D(\theta, \theta_2) \to \infty$ at different rates (θ_1 and θ_2, respectively) as $\theta \downarrow -1$. Hence, (8) is always strictly greater than 1 in this setting.

Example 3. Consider a family that puts positive mass on the left endpoint of its support, which we denote by L. As we send $\theta \to -\infty$, the distribution becomes a point mass at L. The distributions for different θ_1 and θ_2 put different amounts of mass at L, so it is straightforward to discriminate between them. From a technical standpoint, $D(\theta, \theta_1)$ and $D(\theta, \theta_2)$ converge to different finite values as $\theta \to -\infty$. Hence, (8) is always strictly greater than 1 in such settings.

As noted earlier, Theorem 1 establishes under Assumption 2 that the regret tail of an algorithm optimized for $\mathcal{M}^K$, with $\mathcal{M}$ being an exponential family, is truncated Cauchy for every environment in $\mathcal{M}^K$. However, regardless of whether or not Assumption 2 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 (5) in Theorem 1 by taking the difference $\theta_{k(1)} - \theta_{k(2)}$ to be sufficiently small and using the relevant continuity property of the ratio of KL divergences on the right side of (5). 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 $\mathcal{M}$ be an exponential family as in (1). Suppose the algorithm π is optimized for $\mathcal{M}^K$, i.e., Assumption 1 holds. Then for any $\epsilon > 0$, there exists $\delta > 0$ such that for any environment $\nu = (\theta_1, \ldots, \theta_K) \in \mathcal{M}^K$ with $0 < \theta_{k(1)} - \theta_{k(2)} < \delta$,*

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(2)}(T) \geq x)}{\log(x)} \geq -(1+\epsilon)$$

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

3.2. Sketch of Theorem 1 Proof and Further Results

Below we provide a proof sketch in the two-armed bandit setting of a simplified version of Theorem 1. 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 from Assumption 1 and a "one-sided" and more general version of the result in Proposition 5 in Section 3.4.

PROPOSITION 3. *Let $\mathcal{M}$ be an exponential family as in (1). Suppose the algorithm π is optimized for $\mathcal{M}^K$, i.e., Assumption 1 holds. Then for any environment $\nu = (\theta_1, \ldots, \theta_K) \in \mathcal{M}^K$ and any sub-optimal arm k,*

$$\frac{N_k(T)}{\log(T)} \to \frac{1}{D(\theta_k, \theta_*)}$$

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

Proceeding to the proof sketch, let $c \in (0,1)$, and $\nu = (\theta_1, \theta_2) \in \mathcal{M}^2$ such that (without loss of generality) $\theta_1 > \theta_2$, i.e., arm 1 is optimal. Under Assumption 1, we will first obtain:

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > cT)}{\log(T)} \geq - \inf_{\theta\in\Theta\,:\,\theta<\theta_2<\theta_1} \frac{D(\theta,\theta_1)}{D(\theta,\theta_2)}. \tag{9}$$

Then additionally under Assumption 2, we will obtain:

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) > cT)}{\log(T)} = -1. \tag{10}$$

Proof Sketch for (9)-(10) To obtain (9), consider a new environment $\widetilde{\nu} = (\widetilde{\theta}_1, \theta_2)$ with $\widetilde{\theta}_1 < \theta_2$, i.e., arm 1 is now sub-optimal. Performing a change of measure from ν to $\widetilde{\nu}$,

$$\mathbb{P}_{\nu\pi}\left(N_2(T) > cT\right) = \mathbb{E}_{\widetilde{\nu}\pi}\left[\prod_{t=1}^{N_1(T)} \frac{p(X_1(t),\theta_1)}{p(X_1(t),\widetilde{\theta}_1)}\,;\, N_2(T) > cT\right]. \tag{11}$$

Defining $\mathcal{E}_T = \{N_1(T) \approx \log(T)/D(\widetilde{\theta}_1, \theta_2)\}$, we have $\{N_2(T) > cT\} \supset \mathcal{E}_T$ for large T. Under $\widetilde{\nu}$ and on event $\mathcal{E}_T$, the likelihood ratio $\prod_{t=1}^{N_1(T)} p(X_1(t),\theta_1)/p(X_1(t),\widetilde{\theta}_1)$ in (11) can be approximated via:

$$\exp\left(-N_1(T)\frac{1}{N_1(T)}\sum_{t=1}^{N_1(T)} \log\frac{p(X_1(t),\widetilde{\theta}_1)}{p(X_1(t),\theta_1)}\right) \approx \exp\left(-N_1(T)D(\widetilde{\theta}_1,\theta_1)\right) \approx \exp\left(-\frac{D(\widetilde{\theta}_1,\theta_1)}{D(\widetilde{\theta}_1,\theta_2)}\log(T)\right).$$

So for large T, we approximately have:

$$\mathbb{P}_{\nu\pi}\left(N_2(T) > cT\right) \geq \exp\left(-\frac{D(\widetilde{\theta}_1,\theta_1)}{D(\widetilde{\theta}_1,\theta_2)}\log(T)\right) \cdot \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{E}_T). \tag{12}$$

By Proposition 3, $\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{E}_T) \approx 1$. (Under the new environment $\widetilde{\nu}$, π continues behaving like an optimized algorithm, but now with arm 1 as the sub-optimal arm.) Taking log of both sides of (12), dividing by $\log(T)$ and sending $T \to \infty$, and finally optimizing over $\widetilde{\theta}_1$, we obtain (9). Using Lemma 1, the right side of (9) equals -1 under Assumption 2. For optimized algorithms satisfying Assumption 1, Markov's inequality indicates that

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

So under Assumptions 1-2, we obtain (10). □

In light of (8), the "optimal" change of measure (from ν to $\widetilde{\nu}$) to obtain (9) essentially involves sending $\widetilde{\theta}_1 \downarrow \inf\Theta$, which can be quite extreme. For example, under the conditions of Proposition 1, $\inf\Theta = -\infty$ and the optimal change of measure would involve sending the optimal arm 1 mean $A'(\widetilde{\theta}_1) \to -\infty$. 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 $\widehat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{i=1}^{N_k(t)} X_k(i)$ to denote the sample mean of arm k rewards up to time t.)

PROPOSITION 4. *Let $\mathcal{M}$ be an exponential family as in (1). Suppose the algorithm π is optimized for $\mathcal{M}^K$, i.e., Assumption 1 holds. Then for any environment $\nu = (\theta_1, \ldots, \theta_K) \in \mathcal{M}^K$ and any sub-optimal arm k, and any $\epsilon > 0$,*

$$\lim_{T\to\infty} \mathbb{P}_{\nu\pi} \left(|\widehat{\mu}_k(T) - \mu(\theta_k)| \leq \epsilon \mid N_k(T) \geq x \right) = 1$$

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

It is straightforward to extend the proof sketch for (9) and (10) to multi-armed settings. To obtain lower bounds on the distribution tail of the number of plays $N_{k(i)}(T)$ of arm $k(i)$ (the i-th best arm, for $i \geq 2$), we tilt the reward distributions of arms $k(1), \ldots, k(i-1)$ so that their means become less than that of arm $k(i)$. We choose the new environment $\widetilde{\nu}$ with the new arm parameter values, so that arm $k(i)$ 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 $k(1), \ldots, k(i-1)$. Subsequently, each of the tilted parameter values for arms $k(1), \ldots, k(i-1)$ can be optimized separately to yield, for example, (5). We refer the reader to the full proof of Theorem 1 in Section 6.1.

3.3. Upper Bounds on Tail Probabilities

In Theorem 1 from Section 3.1, we developed a lower bound (5) for the distribution tail of the number of plays $N_{k(i)}(T)$ of arm $k(i)$ (the i-th best arm, for $i \geq 2$) by an algorithm optimized for an exponential family class of environments (satisfying Assumption 1). Under Assumption 2, we showed in (6) that the tail exponent for $N_{k(2)}(T)$ is exactly equal to -1. The lower bound part of this result is obtained using (5), Assumption 2 and Lemma 1. The upper bound part of this result

is obtained using Assumption 1 and Markov's inequality, which together imply that for any arm $k(i)$ with $i \geq 2$,

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T) > x)}{\log(x)} \leq -1 \tag{13}$$

uniformly for $x \in [T^\gamma, (1-\gamma)T]$ for $\gamma \in (0,1)$ as $T \to \infty$. (This upper bound also holds for all consistent algorithms satisfying (3).)

However, when Assumption 2 does not hold, then the lower bound in (5) and the upper bound in (13) do not match for arm $k(2)$. Moreover, regardless of whether or not Assumption 2 holds, the lower bound in (5) and the upper bound in (13) do not match for any arm $k(i)$ with $i \geq 3$. As part of Theorem 2, we develop upper bounds for the tail of $N_{k(i)}(T)$ for the KL-UCB algorithm. (See Algorithm 2 and Theorem 1 of Cappé et al. (2013).) These upper bounds exactly match the lower bounds in (5), thereby providing strong evidence that the lower bounds in (5) are tight more generally, regardless of whether or not Assumption 2 holds. The proof of Theorem 2 is given in Section 6.2.

THEOREM 2. *Let $\mathcal{M}$ be an exponential family as in (1). Let the algorithm π be KL-UCB optimized for $\mathcal{M}^K$. Then for any environment $\nu = (\theta_1, \ldots, \theta_K) \in \mathcal{M}^K$ and any $i \geq 2$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T) \geq x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{\theta\in\Theta\,:\,\theta<\theta_{k(i)}} \frac{D(\theta,\theta_{k(j)})}{D(\theta,\theta_{k(i)})} \tag{14}$$

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

From (14), we see that the tail exponents for the distributions of $N_{k(i)}(T)$, $i \geq 3$ are always strictly less than that of $N_{k(2)}(T)$. Indeed, when Assumption 2 holds, Lemma 1 implies that the right side of (14) is exactly $-(i-1)$ for $i \geq 3$, which can be compared with (6). So the second-best arm $k(2)$ determines the exponent of the distribution tail of the aggregate regret; see Remark 1 below. Whenever Assumption 2 is not satisfied (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 (14) is always strictly less than -1 for the second-best arm $k(2)$. 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 $\mathcal{M}^K$, when $\mathcal{M}$ is a discrete exponential family, is less fragile than when $\mathcal{M}$ is a continuous exponential family. However, recall from Corollary 1 that regardless of whether the reward distributions are discrete or continuous, there always exist environments in $\mathcal{M}^K$ 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 $R(T) = \sum_{i=2}^{K} N_{k(i)}(T)\Delta_{k(i)}$ is determined by that of $N_{k(2)}(T)$. This is because for sufficiently large T, and for $x \in [c_1 T^\gamma, c_2 T]$ for some $c_1 > 0$ and $c_2 > 0$ and any $\gamma \in (0,1)$,

$$\begin{aligned}\frac{\log \mathbb{P}_{\nu\pi}\left(N_{k(2)}(T) > x/\Delta_{k(2)}\right)}{\log(x)} &\leq \frac{\log \mathbb{P}_{\nu\pi}(R(T) > x)}{\log(x)} \\ &\leq \frac{\log \sum_{i=2}^{K} \mathbb{P}_{\nu\pi}\left(N_{k(i)}(T) > x/(K\Delta_{k(i)})\right)}{\log(x)} \\ &\leq \frac{\log \mathbb{P}_{\nu\pi}\left(N_{k(2)}(T) > x/(K\Delta_{k(K)})\right)}{\log(x)} + \frac{\log(K)}{\log(x)}.\end{aligned}$$

Sending $T \to \infty$, we conclude that the tail of $R(T)$ is determined by that of $N_{k(2)}(T)$.

We also point out that (14) 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 (6) in Theorem 1. Since we simply relied on logarithmic expected regret and Markov's inequality in Theorem 1 to establish the upper bound part of (6), we could not make conclusions about the regret distribution tail in poly-log(T) regions. Here 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 KL-UCB optimized for environments within classes $\mathcal{M}^K$, for exponential families $\mathcal{M}$. Later in Theorem 4, we develop a result for general UCB algorithms operating in essentially arbitrary environments.

3.4. A New Proof of the Generalized Lai-Robbins Lower Bound

In Corollary 1, we saw that for any algorithm optimized for an exponential family, there always exist environments with very close top arm parameters/means for which the algorithm produces regret tails that are arbitrarily close to being truncated Cauchy. This suggests that an optimized algorithm cannot be further optimized or else the regret tails will become heavier than truncated Cauchy for such environments, which would violate consistency. In this section, we show that this idea can be adapted to provide a new proof of the generalized version (Theorem 3 below, due to Burnetas and Katehakis (1996)) of the Lai-Robbins lower bound for expected regret (from (4)).

Before developing Theorem 3, we first present the following heuristic derivation of the simpler Lai-Robbins lower bound in (4). The proof of Theorem 3 then formalizes and generalizes the key ideas in this heuristic derivation. For simplicity, we consider the two-armed setting.

Proof Sketch for (4) Let π be a consistent algorithm (as in (3)) for environments in $\mathcal{M}^2$, where $\mathcal{M}$ is an exponential family as in (1). Consider a generic environment $\widetilde{\nu} = (\widetilde{\theta}_1, \theta_2)$, with $\widetilde{\theta}_1 < \theta_2$ (arm 1 is sub-optimal, without loss of generality). We claim that $\mathbb{E}_{\widetilde{\nu}\pi}[N_1(T)]$ cannot grow more slowly than $\log(T)/D(\widetilde{\theta}_1, \theta_2)$ as $T \to \infty$. To see why, suppose $(1-2\epsilon)\log(T)/D(\widetilde{\theta}_1, \theta_2)$ is the growth rate, for some $\epsilon \in (0, 1/2)$. Then, with $\mathcal{E}_T = \{N_1(T) \leq (1-\epsilon)\log(T)/D(\widetilde{\theta}_1, \theta_2)\}$, we have $\liminf_{T\to\infty} \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{E}_T) > 0$. Now consider how π behaves in a "hard" environment $\nu = (\theta_1, \theta_2)$, with $\theta_1 > \theta_2$ (arm 1 is optimal) and θ_1 sufficiently close to θ_2 so that $D(\widetilde{\theta}_1, \theta_1)/D(\widetilde{\theta}_1, \theta_2) \leq 1+\epsilon$ (as in Corollary 1). By the change of measure argument from the Theorem 1 proof sketch (see (11)-(12)),

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) \geq T/2)}{\log(T)} \geq -(1-\epsilon)\frac{D(\widetilde{\theta}_1, \theta_1)}{D(\widetilde{\theta}_1, \theta_2)} + \liminf_{T\to\infty} \frac{\log \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{E}_T)}{\log(T)} \geq -(1-\epsilon^2).$$

This violates consistency for the hard environment ν, and so our claim must hold. $\square$

From the above heuristic derivation, we see that the "hardest" environments with very close top arm parameters/means (as found in Corollary 1) determine the minimum growth rate of expected regret in all environments. In every environment, an algorithm must sample sub-optimal arms at least according to the minimum rate in (4) in order to guard against accruing super-logarithmic expected regret if the environment turns out to be one of these hardest cases. Intuitively, in accordance with Corollary 1, the regret tail of an optimized algorithm is already very close to (if not exactly) truncated Cauchy in these hardest environments. So designing the algorithm to be more aggressive overall by exploring less/exploiting more would cause the regret tail to become heavier than truncated Cauchy in such environments, which would violate consistency. (In the above proof sketch for (4), the only assumption used is consistency of the algorithm under consideration. Unlike in the proof sketch of Theorem 1, Proposition 3 is not used. In fact, we later use the ideas from the above proof sketch for (4) to prove Proposition 3; see Proposition 5 below.)

We now formally introduce and prove Theorem 3. This result was first given, with a different proof, in Proposition 1 of Burnetas and Katehakis (1996). Building on the Lai-Robbins lower bound for expected regret, where the model $\mathcal{M}$ is an exponential family, $\mathcal{M}$ is allowed to be an arbitrary collection of distributions (with finite means) in Theorem 3. For such an arbitrary $\mathcal{M}$, an arbitrary distribution P, and $z \in \mathbb{R}$, we define:

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

where $D(P, P')$ denotes the KL divergence between the distributions P and P'. In Theorem 3, the environments $\nu \in \mathcal{M}^K$ 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 a consistent algorithm for the class of environments $\mathcal{M}^K$, i.e., π satisfies (3) for any $a > 0$, environment $\nu \in \mathcal{M}^K$ and sub-optimal arm k. Then for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and any sub-optimal arm k,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}[N_k(T)]}{\log(T)} \geq \frac{1}{D_{inf}(P_k, \mu_*, \mathcal{M})}. \tag{15}$$

Theorem 3 follows immediately from Proposition 5 below and Markov's inequality. When $\mathcal{M}$ is an exponential family model as in (1), (15) simplifies to (4). Moreover, as mentioned earlier, Proposition 5 and Assumption 1 directly imply Proposition 3 when $\mathcal{M}$ is an exponential family. The proof of Proposition 5 is based on the key ideas from the heuristic derivation of the Lai-Robbins lower bound given earlier in this section. In the proof, for any distributions P and P', we use dP/dP' to denote the Radon-Nikodym derivative of the absolutely continuous part of P with respect to P', in accordance with the Lebesgue decomposition of P with respect to P' (see Theorem 6.10 of Rudin (1987) for a precise statement).

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

$$\lim_{T\to\infty} \mathbb{P}_{\nu\pi}\left(\frac{N_k(T)}{\log(T)} \geq \frac{1}{D_{inf}(P_k, \mu_*, \mathcal{M})}\right) = 1. \tag{16}$$

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 (16) is false. Without loss of generality, suppose arm 1 is sub-optimal in $\widetilde{\nu}$, and there exists $\epsilon \in (0,1)$ and a sequence of deterministic times $T_n \uparrow \infty$ such that for all n,

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(\frac{N_1(T_n)}{\log(T_n)} \leq \frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1, \mu_*, \mathcal{M})}\right) \geq \epsilon. \tag{17}$$

Consider another environment $\nu = (P_1, P_2, \ldots, P_K) \in \mathcal{M}^K$, with $\mu(P_1) > \mu_*$ and

$$\frac{D(\widetilde{P}_1, P_1)}{D_{\inf}(\widetilde{P}_1, \mu_*, \mathcal{M})} \leq 1 + \epsilon. \tag{18}$$

Such $P_1 \in \mathcal{M}$ exists or else $D_{\inf}(\widetilde{P}_1, \mu_*, \mathcal{M}) = \infty$ (infimum over an empty set), resulting in (16) being trivially satisfied for $\widetilde{\nu}$. Define the events $\mathcal{A}_n = \{\exists\, k \in \{2, \ldots, K\} : N_k(T_n) \geq T_n/(2K)\}$,

$$\mathcal{B}_n = \left\{\frac{N_1(T_n)}{\log(T_n)} \leq \frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1, \mu_*, \mathcal{M})}\right\}, \quad \mathcal{C}_n = \left\{\frac{1}{N_1(T_n)} \sum_{t=1}^{N_1(T_n)} \log \frac{d\widetilde{P}_1}{dP_1}(X_1(t)) \leq D(\widetilde{P}_1, P_1) + \delta\right\},$$

with $\delta > 0$, and note that $\mathcal{A}_n \supset \mathcal{B}_n$ for large n. Performing a change of measure from ν to $\widetilde{\nu}$,

$$\begin{aligned}
\mathbb{P}_{\nu\pi}(\mathcal{A}_n) &\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\prod_{t=1}^{N_1(T_n)} \frac{dP_1}{d\widetilde{P}_1}(X_1(t))\,;\,\mathcal{A}_n\right] \qquad (19)\\
&\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\exp\left(-N_1(T_n)\frac{1}{N_1(T_n)}\sum_{t=1}^{N_1(T_n)} \log\frac{d\widetilde{P}_1}{dP_1}(X_1(t))\right);\mathcal{B}_n,\mathcal{C}_n\right]\\
&\geq \exp\left(-\Big(D(\widetilde{P}_1,P_1)+\delta\Big)\frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1,\mu_*,\mathcal{M})}\log(T_n)\right)\cdot\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_n,\mathcal{C}_n),
\end{aligned}$$

with an inequality in (19), since P_1 may not be absolutely continuous with respect to $\widetilde{P}_1$. Applying a union bound to $\mathcal{A}_n$, taking logs and dividing by $\log(T_n)$, we have for some $k' \in \{2,\ldots,K\}$,

$$\frac{\log\Big((K-1)\cdot\mathbb{P}_{\nu\pi}\big(N_{k'}(T_n)\geq T_n/(2K)\big)\Big)}{\log(T_n)} \geq -\Big(D(\widetilde{P}_1,P_1)+\delta\Big)\frac{1-\epsilon}{D_{\inf}(\widetilde{P}_1,\mu_*,\mathcal{M})} + \frac{\log\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_n,\mathcal{C}_n)}{\log(T_n)}.$$

By Lemma 2 (in Appendix D) and the WLLN for sample means, $\lim_{n\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{C}_n)=1$. So from (17), $\liminf_{n\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_n,\mathcal{C}_n)\geq\epsilon$. Upon sending $\delta\downarrow 0$ and using (18),

$$\liminf_{n\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_{k'}(T_n)\geq T_n/(2K))}{\log(T_n)} \geq -(1-\epsilon)\cdot\frac{D(\widetilde{P}_1,P_1)}{D_{\inf}(\widetilde{P}_1,\mu_*,\mathcal{M})} \geq -(1-\epsilon^2).$$

Since $\epsilon\in(0,1)$, this violates the consistency of π, and thus (17) cannot be true. □

Remark 2. To the best of our knowledge, there are three alternative proofs of Theorem 3 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), ours uses the heaviness of the regret distribution tail to reason about what the minimum growth rate of expected regret must be. Our proof highlights the fact that environments with very close arm means are the hardest cases, with regret tails that are arbitrarily close to being truncated Cauchy. As seen using our change of measure argument, in order to prevent the regret tails for these environments from becoming heavier than truncated Cauchy, expected regret must grow at a minimum rate in all environments.

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 rely on general-purpose information-theoretic results, while we argue directly from first principles with our change-of-measure approach. Additionally, Kaufmann and Cappé (2016) (in Theorem 21) develop a similar asymptotic lower bound for expected regret using sophisticated tools tailored for best-arm identification settings. 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 and Cappé (2016) only applies when there is a unique optimal arm.

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 on the KL-UCB algorithm (see Algorithm 2 of Cappé et al. (2013)) throughout Sections 4.1-4.3. Unless specified otherwise, our theory will be developed for KL-UCB optimized for any chosen exponential family class of environments $\mathcal{M}^K$, and operating in an environment $\nu \notin \mathcal{M}^K$. In such settings, we will use d to denote the KL divergence function corresponding to the chosen exponential family $\mathcal{M}$ (with no relation to the true environment ν). For convenience, we will use the form of d that takes mean values as inputs, i.e., $d : I \times I \to [0, \infty)$, with I being an open interval (possibly infinite) corresponding to the means of $\mathcal{M}$. To avoid pathological situations, we will always assume that I contains the (long-run) mean of every arm in the true environment ν. 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 Theorem 4, which is a characterization of the regret tail of (possibly) mis-specified KL-UCB operating in an environment $\nu = (P_1, \ldots, P_K)$, where arm k yields independent rewards from some arbitrary distribution P_k. For distribution P_k, we use P_k^z to denote its exponentially tilted version having mean z. For arm $k(i)$ with $i \geq 2$, we can see on the right side of (20), the key role played by the exponentially tilted reward distributions $P_{k(j)}^z$, $1 \leq j \leq i-1$. We recall that such tilting is also key to Theorems 1 and 2. (Note that the KL divergence $D(P_{k(j)}^z, P_{k(j)}) = \Lambda_{k(j)}^*(z)$, where $\Lambda_{k(j)}^*$ is the convex conjugate of the CGF $\Lambda_{k(j)}$ of $P_{k(j)}$.) We can compare the right side of (20) in Theorem 4 to the right side of (14) in Theorem 2. The same structure is present in both cases, namely, the ratios involve KL divergence between different exponential tilts of the arm reward distributions. In well-specified settings, Theorem 2 and Theorem 4 are the same result. In mis-specified settings, which is covered by Theorem 4, the KL divergence D in the numerator does not match the KL divergence d in the denominator.

The proof of Theorem 4 is given in Appendix E. The proof uses a SLLN (Proposition 6) and a tail probability lower bound (Theorem 5) for the regret of (possibly) mis-specified KL-UCB, which are deferred to Section 4.2. These supporting results are developed for more general (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.

THEOREM 4. *Let π be KL-UCB with divergence d. Let the environment $\nu = (P_1, \ldots, P_K)$, where arm k yields independent rewards from an arbitrary distribution P_k, with CGF Λ_k such that $\Lambda_k(\lambda) < \infty$ for λ in a neighborhood of zero. Then for any $i \geq 2$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T) \geq x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{z<\mu(P_{k(i)})} \frac{D(P_{k(j)}^z, P_{k(j)})}{d(z, \mu(P_{k(i)}))} \tag{20}$$

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. We verify (21) through numerical simulations in Figure 1 in Section 7.

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 $\sigma_0^2 > 0$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(2)}(T) \geq x)}{\log(x)} = -\frac{\sigma^2}{\sigma_0^2} \tag{21}$$

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 $a \in (\sigma^2/\sigma_0^2, 1]$,

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

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. Lower Bound 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 (and also in Section 4.1 in settings where the marginal reward distributions are mis-specified). Proposition 6 is a SLLN for the regret of KL-UCB operating in an environment with general (possibly non-iid) reward processes that satisfy Assumptions 3-4 below. (For all results depending on Proposition 6, we actually only need the weak version of Proposition 6 (a WLLN) with convergence in $\mathbb{P}_{\nu\pi}$-probability.) Then in Theorem 5, we apply our change-of-measure argument to establish lower bounds for the regret tail of KL-UCB when operating in an environment with such reward processes.

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

$$\Lambda_k^n(\lambda) = \frac{1}{n} \log \mathbb{E}\left[\exp\left(\lambda \cdot \sum_{i=1}^n X_k(i)\right)\right], \quad \lambda \in \mathbb{R}.$$

We will assume the following for each arm k.

ASSUMPTION 3. *The limit $\Lambda_k(\lambda) = \lim_{n\to\infty} \Lambda_k^n(\lambda)$ exists for each $\lambda \in \mathbb{R}$ (possibly infinite), and $0 \in \mathcal{L}_k = int\{\lambda \in \mathbb{R} : \Lambda_k(\lambda) < \infty\}$.*

ASSUMPTION 4. *$\Lambda_k(\cdot)$ is differentiable throughout $\mathcal{L}_k$, and $\lim_{m\to\infty} |\Lambda_k'(\lambda_m)| = \infty$ for any sequence $\lambda_m \in \mathcal{L}_k$ converging to a boundary point of $\mathcal{L}_k$.*

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 3, we refer to the limit Λ_k as the *limiting CGF* (for arm k). In the context of Assumption 4, $\Lambda_k'(0)$, the derivative of limiting CGF evaluated at zero, is the long-run mean reward for arm k. Indeed, by the Gärtner-Ellis Theorem and the Borel-Cantelli Lemma,

$$\frac{1}{n}\sum_{i=1}^{n} X_k(i) \to \Lambda_k'(0) \tag{22}$$

almost surely as $n \to \infty$ for each arm k. The optimal arm $k(1)$ is such that $\Lambda_{k(1)}'(0) = \max_k \Lambda_k'(0)$.

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 3 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 rewards from different arms are independent.

We now state Proposition 6 and Theorem 5. Their proofs can be found in Appendix F. In the context of Theorem 5, for an arm k, Λ_k^* is the convex conjugate of the limiting CGF Λ_k. We recall that the convex conjugate of the limiting CGF is the *rate function* in the Gärtner-Ellis Theorem.

PROPOSITION 6. *Let π be KL-UCB with divergence d. Let the K-armed environment ν yield rewards for each arm that evolve according to any process satisfying Assumptions 3-4. Then for any $i \geq 2$, $\mathbb{P}_{\nu\pi}$-almost surely as $T \to \infty$,*

$$\frac{N_{k(i)}(T)}{\log(T)} \to \frac{1}{d(\Lambda_{k(i)}'(0), \Lambda_{k(1)}'(0))}.$$

THEOREM 5. *Let π be KL-UCB with divergence d. Let the K-armed environment ν yield rewards for each arm that evolve according to any process satisfying Assumptions 3-4. Then for any $i \geq 2$,*

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T) \geq x)}{\log(x)} \geq -\sum_{j=1}^{i-1} \inf_{z < \Lambda_{k(i)}'(0)} \frac{\Lambda_{k(j)}^*(z)}{d(z, \Lambda_{k(i)}'(0))}$$

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

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 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

(designed for iid rewards) applied to bandits with rewards evolving as Gaussian AR(1) processes. Specifically, for each arm k, we assume the rewards evolve as an AR(1) process:

$$X_k(t) = \alpha_k + \beta_k X_k(t-1) + W_k(t), \tag{23}$$

where the $\beta_k \in (0,1)$ and the $W_k(t)$ are iid $N(0,\sigma_k^2)$. The equilibrium distribution for arm k is then $N(\alpha_k/(1-\beta_k), \sigma_k^2/(1-\beta_k^2))$. 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 (22)) for arm k is $\Lambda_k'(0) = \alpha_k/(1-\beta_k)$. The proof of Corollary 3 follows from a straightforward verification of Assumptions 3-4, which is omitted, and then a direct application of Theorem 5.

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 (23)),*

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(2)}(T) \geq x)}{\log(x)} \geq -\frac{\sigma^2}{\sigma_{k(1)}^2}(1-\beta_{k(1)})^2$$

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

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,

$$\liminf_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(2)}(T) \geq x)}{\log(x)} \geq -\frac{1-\beta_0}{1+\beta_0}, \tag{24}$$

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

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

We verify (24) through numerical simulations in Figure 2 in Section 7. The simulations suggest that the lower bound in (24) is tight.

In some settings, there is essentially no risk of mis-specifying the marginal reward distributions. This is the case for Bernoulli reward distributions as well as other discrete distributions with finite support. However, it is often still the case that the serial dependence structure of the rewards can be mis-specified, which can then cause the performance of optimized algorithms to deteriorate. 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 k, we assume that the rewards evolve as an irreducible Markov chain on a common, finite state space $S \subset \mathbb{R}$, with transition matrix Q_k. For any $\lambda \in \mathbb{R}$, we use $\rho_k(\lambda)$ to denote the Perron-Frobenius eigenvalue of the tilted matrix

$$Q_k^\lambda = (\exp(\lambda \cdot y) Q_k(x,y),\ x,y \in S). \tag{25}$$

In the context of Assumptions 3-4, $\Lambda_k(\lambda) = \log \rho_k(\lambda)$ (recall that Λ_k is the limiting CGF). (Note that the convex conjugate Λ_k^* of Λ_k plays the same role in Corollary 4 as it does in Theorem 5.) 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 (22)) for arm k is $\Lambda_k'(0) = \rho_k'(0)/\rho_k(0)$. 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 (25). This provides technical convenience, and allows us to use Chernoff 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 Q on S satisfies the *Doeblin Condition* if we have $Q(x, s_{\min}) > 0$ for each $x \neq s_{\min}$, and $Q(x, s_{\max}) > 0$ for each $x \neq s_{\max}$.

The lower bound part of Corollary 4 follows from a straightforward verification of Assumptions 3-4, which is omitted, and then a direct application of Theorem 5. 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 (52) and (56)), a Chernoff bound for additive functionals of finite-state Markov chains established in Theorem 1 of Moulos and Anantharam (2019).

COROLLARY 4. *Let π be KL-UCB with divergence d. 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 Λ_k as defined above for each arm k), and suppose that the transition matrix for each arm satisfies the Doeblin Condition. Then for any $i \geq 2$,*

$$\lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T) \geq x)}{\log(x)} = -\sum_{j=1}^{i-1} \inf_{z < \Lambda'_{k(i)}(0)} \frac{\Lambda^*_{k(j)}(z)}{d(z, \Lambda'_{k(i)}(0))} \tag{26}$$

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

EXAMPLE 4. For the state space $S = \{0,1\}$ (binary rewards), we can examine some numerical values for the right side of (26). Here, we take $d(z,z')$ 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 $k(1)$ evolves according to a transition matrix of the form:

$$Q_{k(1)} = \begin{bmatrix} 1-p & p \\ 0.1 & 0.9 \end{bmatrix}. \tag{27}$$

Suppose also that the gap between the equilibrium means of the top two arms ($k(1)$ and $k(2)$) is $\Delta > 0$. In Table 1 below, we provide numerical values for the right side of (26) for the case $i = 2$ and different values of p and Δ. As p becomes smaller relative to 0.9, the autocorrelation in the rewards for arm $k(1)$ 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 p and Δ) to obtain regret tails that are heavier than truncated Cauchy (the right side of (26) is greater than -1).

	Δ								
p	0.12	0.11	0.10	0.09	0.08	0.07	0.06	0.05	0.04
0.9	-1.52	-1.48	-1.43	-1.39	-1.34	-1.30	-1.26	-1.21	-1.17
0.8	-1.10	-1.07	-1.03	-1.00	-0.97	-0.94	-0.91	-0.88	-0.85
0.7	-0.86	-0.83	-0.81	-0.78	-0.76	-0.74	-0.71	-0.69	-0.67
0.6	-0.69	-0.67	-0.65	-0.63	-0.61	-0.59	-0.57	-0.56	-0.54

Table 1 For arm rewards evolving as Markov chains on the state space $S = \{0, 1\}$, and with d being the Bernoulli KL divergence, we provide numerical values for the right side of (26). Here, $i = 2$, and we consider different values of p (as used in the best arm's transition matrix $Q_{k(1)}$ in (27)) and Δ (the gap between the equilibrium means of the two best arms, $k(1)$ and $k(2)$).

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 $\mathcal{M}$ be an exponential family as in (1). Suppose the algorithm π is optimized for $\mathcal{M}^K$, i.e., Assumption 1 holds. Suppose also that Assumption 2 holds for $\mathcal{M}$. Then for any environment $\nu \in \mathcal{M}^K$, and any $\delta > 0$ and $\delta' \in (0, \delta)$,*

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

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. In Section 5.1, we first describe a simple way to construct a UCB algorithm with logarithmic expected regret across all environments from a general class $\mathcal{M}^K$. We then describe how to modify the algorithm to ensure the regret tail exponent is at the level of $-(1+b)$ (or less) for any desired $b>0$, uniformly across $\mathcal{M}^K$. In Section 5.2, we show that the modification provides some protection against mis-specification of the arm reward distributions in iid reward settings. In Section 5.3, we show that this type of modification also provides some protection against Markovian departures from independence of the arm rewards. In both cases, the modification to ensure a lighter regret tail, uniformly for all environments in some class $\mathcal{M}^K$, has the added benefit of ensuring logarithmic expected regret for all environments in some strictly larger class. Our analysis also provides an explicit trade-off between the amount of UCB exploration and the resulting heaviness of the regret distributional tail, as discussed in Remark 3.

5.1. A Simple Approach to Obtain Lighter Regret Tails

In this section, we show how to ensure the regret distribution tail exponent is at the level of $-(1+b)$ (or less) for some desired $b>0$, across all environments from a general class $\mathcal{M}^K$. For simplicity, we begin by showing this when $\mathcal{M}$ is an exponential family, as in (1). Let $d(z,z')$ denote the KL divergence between distributions with means z and z' in the exponential family $\mathcal{M}$. (In this section, as well as in Sections 5.2 and 5.3, the divergence d will always be chosen to correspond to the KL divergence between distributions in some stated exponential family. We will always assume that d takes mean values as inputs.) If we take the algorithm π to be KL-UCB using the divergence $d/(1+b)$, then by a direct adaptation of Theorem 4, we have for all environments $\nu=(\theta_1,\ldots,\theta_K)\in\mathcal{M}^K$, and any $i\geq 2$,

$$\begin{aligned}\lim_{T\to\infty}\frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T)>x)}{\log(x)} &= -(1+b)\sum_{j=1}^{i-1}\inf_{z<\mu(\theta_{k(i)})}\frac{d(z,\mu(\theta_{k(j)}))}{d(z,\mu(\theta_{k(i)}))} \\ &\leq -(1+b)(i-1)\end{aligned} \tag{28}$$

uniformly for $x\in[\log^{1+\gamma}(T),(1-\gamma)T]$ for any $\gamma\in(0,1)$ as $T\to\infty$. See Figure 4 in Section 7 for a numerical demonstration of this result when $\mathcal{M}$ is a Gaussian family.

We now consider the case where $\mathcal{M}$ is a general family of distributions. We begin by describing a simple UCB algorithm (inspired by the discussion in Section 6.1 of Cappé et al. (2013) and Chapter

2.2 of Bubeck and Cesa-Bianchi (2012)) to obtain logarithmic expected regret for all environments in the general class $\mathcal{M}^K$. For any distribution $P \in \mathcal{M}$, let Λ_P denote its CGF, which we always assume to be finite on $\mathbb{R}$ for simplicity. The algorithm is simple—we choose an exponential family $\widehat{\mathcal{M}}$ (as in (1)) satisfying the following two properties. First, with I denoting the set of means of the distributions in $\widehat{\mathcal{M}}$, we require $\mu(P) \in I$ for all $P \in \mathcal{M}$. Second, with ψ_z denoting the CGF of the distribution in $\widehat{\mathcal{M}}$ with mean z, and for all $P \in \mathcal{M}$, we require the following domination property:

$$\Lambda_P(\lambda) \le \psi_{\mu(P)}(\lambda), \quad \forall\, \lambda \in \mathbb{R}. \tag{29}$$

The algorithm is then simply KL-UCB designed for $\widehat{\mathcal{M}}$. Since the CGF's of $\widehat{\mathcal{M}}$ dominate those of $\mathcal{M}$ as in (29), we can ensure logarithmic expected regret across all environments in $\mathcal{M}^K$. Below, we provide two examples of a general family $\mathcal{M}$ with a dominating exponential family $\widehat{\mathcal{M}}$.

Example 5. Let $\mathcal{M}$ be the collection of all sub-Gaussian distributions with variance proxy σ^2. (We say Z is sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \le e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$.) Then (29) is satisfied by choosing $\widehat{\mathcal{M}}$ to be the Gaussian family with variance σ^2.

Example 6. Let $\mathcal{M}$ be the family of all distributions supported on $[0,1]$. Then (29) is satisfied by choosing $\widehat{\mathcal{M}}$ to be the Bernoulli family.

Let π be KL-UCB using divergence d corresponding to the KL divergence for $\widehat{\mathcal{M}}$. Then, by using a straightforward adaptation of the proof of Theorem 10.6 in book Lattimore and Szepesvári (2020), together with Chernoff bounds for distributions in $\mathcal{M}$ based on the domination condition (29), we have for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$, and any $i \ge 2$,

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}\left[N_{k(i)}(T)\right]}{\log(T)} \le \frac{1}{d(\mu(P_{k(i)}), \mu(P_{k(1)}))}. \tag{30}$$

We now aim for a regret tail exponent of $-(1+b)$ (or less) uniformly across all environments from $\mathcal{M}^K$, for some desired $b > 0$. Corollary 6, which follows from a direct application of Theorem 4, describes how to do so.

Corollary 6. *For $b > 0$, let π be KL-UCB using divergence $d/(1+b)$, with d being the KL divergence for an exponential family $\widehat{\mathcal{M}}$, whose CGF's ψ_z dominate those of a general family $\mathcal{M}$, as in (29). Then for any environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$, and any $i \ge 2$,*

$$\begin{aligned} \lim_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}(N_{k(i)}(T) > x)}{\log(x)} &= -(1+b) \sum_{j=1}^{i-1} \inf_{z < \mu(P_{k(i)})} \frac{D(P^z_{k(j)}, P_{k(j)})}{d(z, \mu(P_{k(i)}))} \\ &\le -(1+b)(i-1) \end{aligned} \tag{31}$$

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

The $-(1+b)(i-1)$ upper bound on (31) is due to the fact that the ratios in the infima of (31) are at least 1. (For a distribution P, recall that P^z is the exponential tilt with mean z.) This is because for any $P \in \mathcal{M}$, and any $z < \mu(P)$,

$$\begin{aligned} D(P^z, P) &= \sup_{\lambda \in \mathbb{R}} \{\lambda z - \Lambda_P(\lambda)\} \\ &\geq \sup_{\lambda \in \mathbb{R}} \{\lambda z - \psi_{\mu(P)}(\lambda)\} = d(z, \mu(P)), \end{aligned}$$

where the inequality follows from (29). Note that if $\mathcal{M}$ is an exponential family, and $\widehat{\mathcal{M}}$ is chosen equal to $\mathcal{M}$, then (31) coincides with (28).

REMARK 3. From (28) and (31), 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 divergence function $d/(1+b)$ instead of d 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 (28) and (31). 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.

5.2. Robustness to Mis-specified Reward Distribution

In this section, we show how designing for a lighter regret distribution tail provides robustness to mis-specification of reward distributions. Following the setup of Section 5.1, let the model $\mathcal{M}$ be a general family of reward distributions. Let $\widehat{\mathcal{M}}$ be an exponential family whose CGF's ψ_z dominate those of $\mathcal{M}$ in the sense of (29) (where $\psi_z(\lambda)$, $\lambda \in \mathbb{R}$ denotes the CGF of the particular distribution in $\widehat{\mathcal{M}}$ with mean z). In this section, we take d to be the KL divergence for $\widehat{\mathcal{M}}$. We saw in Section 5.1 that for environments in $\mathcal{M}^K$, using KL-UCB with the divergence d results in regret distribution tails that are lighter than, but possibly close to being truncated Cauchy.

However, when the arm reward distributions are mis-specified (i.e., $\mathcal{M}$ is mis-specified), the regret performance can easily deteriorate. For example, as we saw in Section 4.1 via Theorem 4 and Corollary 2, the regret tail is heavier than truncated Cauchy if the variance for Gaussian (or sub-Gaussian) bandits is even slightly under-specified, resulting in expected regret that grows as a power of the time horizon T. To alleviate such issues, following the procedure from Section 5.1, we aim for a lighter regret tail with an exponent of $-(1+b)$ (or less) for some desired $b > 0$, uniformly across all environments in $\mathcal{M}^K$. As before, we do so by using KL-UCB with divergence

$d/(1+b)$. We will see in Corollary 7 below that this provides some protection against distributional mis-specification of the arm rewards. In particular, we can maintain logarithmic expected regret for environments from the enlarged class $\mathcal{M}_b^K$, which is defined in (34) below (after we introduce a few technical notions).

Regarding the divergence $d/(1+b)$, it turns out that $z \mapsto d(z,z')/(1+b)$ is the convex conjugate of the re-scaled CGF:

$$\Psi_{z'}^b(\lambda) = \frac{\psi_{z'}((1+b)\lambda)}{1+b}, \quad \lambda \in \mathbb{R}. \tag{32}$$

To see this, note that

$$\begin{aligned} \frac{d(z,z')}{1+b} &= \frac{1}{1+b} \cdot \sup_{\lambda \in \mathbb{R}} \{\lambda z - \psi_{z'}(\lambda)\} \\ &= \sup_{\lambda \in \mathbb{R}} \{\lambda z - \Psi_{z'}^b(\lambda)\}. \end{aligned} \tag{33}$$

Using the re-scaled CGF's Ψ_z^b, we define the enlarged model:

$$\mathcal{M}_b = \left\{P : \mu(P) \in I,\, \Lambda_P(\lambda) \le \Psi_{\mu(P)}^b(\lambda)\ \forall\, \lambda \in \mathbb{R}\right\}, \tag{34}$$

where (as in Section 5.1) I is the set of means of the distributions in $\widehat{\mathcal{M}}$. It is straightforward to see that $\mathcal{M}$ is a strict subset of $\mathcal{M}_b$. This is because for any (non-degenerate) $P \in \mathcal{M}$,

$$\Lambda_P(\lambda) < \Lambda_P((1+b)\cdot\lambda)/(1+b) \le \Psi_{\mu(P)}^b(\lambda), \quad \forall\, \lambda \neq 0,$$

where the strict inequality is due to Jensen's inequality, and the non-strict inequality is due to (29) and (32). An example of $\mathcal{M}$, $\widehat{\mathcal{M}}$ and $\mathcal{M}_b$ is the following.

Example 7. Let $\mathcal{M}$ be the collection of all sub-Gaussian distributions with variance proxy σ^2. By choosing $\widehat{\mathcal{M}}$ to be the Gaussian family with variance σ^2, (29) is satisfied. Then $\mathcal{M}_b$ is the collection of all sub-Gaussian distributions with variance proxy $\sigma^2(1+b)$.

We have Corollary 7 below, which can be obtained by using a straightforward adaptation of the proof of Theorem 10.6 in book Lattimore and Szepesvári (2020), together with Chernoff bounds for distributions in $\mathcal{M}_b$ based on the domination condition in its definition in (34). By aiming for a regret tail exponent of $-(1+b)$ (or less) uniformly across all environments from $\mathcal{M}^K$, we have the added benefit of ensuring logarithmic expected regret for all environments in the enlarged class $\mathcal{M}_b^K$. To ensure logarithmic expected regret for the enlarged class $\mathcal{M}_b^K$, it is generally necessary to aim for a lighter regret tail uniformly across the smaller class $\mathcal{M}^K$, as we discussed earlier; see Theorem 4 and Corollary 2.

COROLLARY 7. *For $b > 0$, let π be KL-UCB using divergence $d/(1+b)$, with d being the KL divergence for the exponential family $\widehat{\mathcal{M}}$. Suppose the environment $\nu = (P_1, \ldots, P_K) \in \mathcal{M}_b^K$, as defined in (34) (with the CGF's of $\widehat{\mathcal{M}}$ denoted by ψ_z). Then for any $i \geq 2$,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}\left[N_{k(i)}(T)\right]}{\log(T)} \leq \frac{1+b}{d(\mu(P_{k(i)}), \mu(P_{k(1)}))}.$$

5.3. Robustness to Mis-specified Reward Dependence Structure

Building upon the previous section, here we show that designing for a lighter regret distribution tail provides robustness to mis-specification of the serial dependence structure of rewards. We consider settings where the rewards for each arm evolve according to an irreducible Markov chain on a common finite state space $S \subset \mathbb{R}$. We consider KL-UCB algorithms (and modifications thereof) designed for iid rewards from distributions in an exponential family $\mathcal{M}$, with support S and densities of the form in (1) (with respect to counting measure). In this section, we use $\psi_z(\lambda)$, $\lambda \in \mathbb{R}$ to denote the CGF of the particular distribution in $\mathcal{M}$ with mean z, and we use d to denote the KL divergence for $\mathcal{M}$. As we saw earlier via Theorem 2, for environments in $\mathcal{M}^K$ (yielding iid rewards), using KL-UCB with divergence d results in regret distribution tails that are slightly lighter than truncated Cauchy.

However, when the rewards for each arm are autocorrelated, e.g., evolving according to a Markov chain, the regret performance can easily deteriorate. As we saw in Section 4.3 via Corollary 4, particularly via Example 4, the regret tail is often heavier than truncated Cauchy in such settings, resulting in expected regret that grows as a power of the time horizon T. To alleviate such issues, following the procedure from Section 5.1, we aim for a lighter regret tail with an exponent of $-(1+b)$ (or less) for some desired $b > 0$, for all environments in $\mathcal{M}^K$. As before, we do so by using KL-UCB with the divergence $d/(1+b)$. We will see in Corollary 8 below that this provides some 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 the set $\widetilde{\mathcal{M}}_b$, which is defined in (35) below (after we introduce a few technical notions).

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). For $Q \in \mathcal{S}_{|S|}$, similar to (25), we use $\rho_Q(\lambda)$ to denote the Perron-Frobenius eigenvalue of the tilted matrix $Q^\lambda = (\exp(\lambda \cdot y)Q(x,y),\ x, y \in S)$. We use $\mu(Q)$ denote the equilibrium mean corresponding to chain with transition matrix Q, i.e., $\mu(Q) = \rho_Q'(0)/\rho_Q(0)$. For a desired $b > 0$, we consider the re-scaled CGF $\Psi_z^b(\lambda) = \psi_z((1+b)\lambda)/(1+b)$, as in (32). (Recall that $z \mapsto d(z,z')/(1+b)$ is the convex conjugate of $\Psi_{z'}^b$, as seen in (33).) We define

$$\widetilde{\mathcal{M}}_b = \left\{Q \in \mathcal{S}_{|S|} \,:\, \log \rho_Q(\lambda) \leq \Psi_{\mu(Q)}^b(\lambda)\ \ \forall \lambda \in \mathbb{R}\right\}. \tag{35}$$

It is straightforward to see that the exponential family $\mathcal{M}$ is equivalent to a strict subset of the collection of transition matrices in $\widetilde{\mathcal{M}}_b$ with identical rows. In Example 8 below, we examine the degree to which $\widetilde{\mathcal{M}}_b$ is "larger" than $\mathcal{M}$ for the case $S = \{0, 1\}$.

EXAMPLE 8. Let the state space $S = \{0, 1\}$, and let $\mathcal{M}$ be the Bernoulli family of distributions, with ψ_z denoting the CGF of the Bernoulli distribution with mean z. Consider transition matrices on S of the form:

$$Q = \begin{bmatrix} 1-p & p \\ 1-p' & p' \end{bmatrix}. \tag{36}$$

The more positive the difference $p' - p$, the more positive the autocorrelation between the rewards. In Table 2 below, for different values of $b > 0$, we examine how positive the difference $p' - p$ can be in order for Q to still belong in $\widetilde{\mathcal{M}}_b$ (as defined in (35)), 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 $p' - p$	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 $p, p' \in [0.05, 0.95]$, we give the maximum allowed difference $p' - p$ that ensures the transition matrix Q in (36) belongs in $\widetilde{\mathcal{M}}_b$, as in (35) (and (32)).

We have Corollary 8 below, which (like for Corollary 7) can be obtained by using a straightforward adaptation of the proof of Theorem 10.6 in book Lattimore and Szepesvári (2020), together with the Chernoff bound for finite state space Markov chains in Theorem 1 of Moulos and Anantharam (2019). By aiming for a regret tail exponent of $-(1+b)$ (or less) uniformly across all environments from $\mathcal{M}^K$ (iid environments), we can ensure logarithmic expected regret for Markovian departures from $\mathcal{M}^K$ that lie within $\widetilde{\mathcal{M}}_b$. Again, to ensure logarithmic expected regret in such settings, it is generally necessary to aim for a lighter regret tail, as we discussed earlier; see Corollary 4 and Example 4.

COROLLARY 8. *For $b > 0$, let π be KL-UCB using divergence $d/(1+b)$, with d being the KL divergence for an exponential family $\mathcal{M}$ that is supported on a finite set $\mathcal{S} \subset \mathbb{R}$. For the K-armed environment ν, suppose arm k yields rewards that evolve according to a Markov chain on S with a transition matrix $Q_k \in \widetilde{\mathcal{M}}_b$, as defined in (35) (with the CGF's of $\mathcal{M}$ denoted by ψ_z). Then for any $i \geq 2$,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\nu\pi}\left[N_{k(i)}(T)\right]}{\log(T)} \leq \frac{1+b}{d(\mu(Q_{k(i)}), \mu(Q_{k(1)}))}.$$

REMARK 4. For general reward processes satisfying Assumptions 3-4, 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 5 (unlike in Theorem 4 and Corollary 4). 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 guarantees for our procedure to lighten the regret tail that is described in this section. Nevertheless, our simulations in Figure 5 in Section 7 suggest that we can still ensure the regret tail is lighter to a desired degree using our procedure. (The lower bound in Theorem 5 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 $\theta_1 > \theta_2 > \cdots > \theta_K$ (i.e., $k(i) = i$ for all $1 \le i \le K$). We first show (5) and (6) for the sub-optimal arm $k = 2$. Consider the alternative environment $\widetilde{\nu} = (\widetilde{\theta}_1, \theta_2, \ldots, \theta_K)$, where $\widetilde{\theta}_1 < \theta_2$ and $\theta_2, \ldots, \theta_K$ are the same parameter values from the environment ν. We will consider different values for $\widetilde{\theta}_1$ later in the proof. Let $\delta > 0$, and define the events:

$$\mathcal{B}_T = \left\{ \left| \widehat{\mu}_1(T) - \mu(\widetilde{\theta}_1) \right| \le \delta \right\}, \qquad \mathcal{C}_T = \left\{ \frac{N_1(T)}{\log(T)} \le \frac{1+\delta}{D(\widetilde{\theta}_1, \theta_2)} \right\} \cap \left\{ \frac{N_j(T)}{\log(T)} \le \frac{1+\delta}{D(\theta_j, \theta_2)} \quad \forall\, j \ge 3 \right\},$$

where $\widehat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{i=1}^{N_k(t)} X_k(i)$. We then have

$$\begin{aligned}
&\mathbb{P}_{\nu\pi}(N_2(T) \ge (1-\gamma)T) \ge \mathbb{P}_{\nu\pi}(\mathcal{B}_T, \mathcal{C}_T) && (37)\\
&= \mathbb{E}_{\widetilde{\nu}\pi}\left[\exp\left(\sum_{t=1}^{N_1(T)} X_1(t) \cdot (\theta_1 - \widetilde{\theta}_1) + \left(A(\widetilde{\theta}_1) - A(\theta_1) \right) N_1(T) \right); \mathcal{B}_T, \mathcal{C}_T \right] && (38)\\
&\ge \mathbb{E}_{\widetilde{\nu}\pi}\left[\exp\left(\left((\mu(\widetilde{\theta}_1) - \delta) \cdot (\theta_1 - \widetilde{\theta}_1) + A(\widetilde{\theta}_1) - A(\theta_1) \right) N_1(T) \right); \mathcal{B}_T, \mathcal{C}_T \right] && (39)\\
&= \mathbb{E}_{\widetilde{\nu}\pi}\left[\exp\left(- \left(D(\widetilde{\theta}_1, \theta_1) + \delta(\theta_1 - \widetilde{\theta}_1) \right) N_1(T) \right); \mathcal{B}_T, \mathcal{C}_T \right] && (40)\\
&\ge \exp\left(- \left(D(\widetilde{\theta}_1, \theta_1) + \delta(\theta_1 - \widetilde{\theta}_1) \right) \frac{1+\delta}{D(\widetilde{\theta}_1, \theta_2)} \log(T) \right) \cdot \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_T, \mathcal{C}_T). && (41)
\end{aligned}$$

In (38), we have introduced the likelihood ratio for arm 1 in changing the environment from ν to $\widetilde{\nu}$. In (39), we use the event $\mathcal{B}_T$. In (40), we use the identities $\mu(\theta) = A'(\theta)$ and (2). In (41), we use the event $\mathcal{C}_T$. Taking log and dividing by $\log(T)$ on the left side of (37) and the right side of (41), we have

$$\frac{\log \mathbb{P}_{\nu\pi}(N_2(T) \ge (1-\gamma)T)}{\log(T)} \ge - \left(D(\widetilde{\theta}_1, \theta_1) + \delta(\theta_1 - \widetilde{\theta}_1) \right) \frac{1+\delta}{D(\widetilde{\theta}_1, \theta_2)} + \frac{\log \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_T, \mathcal{C}_T)}{\log(T)}. \qquad (42)$$

Using the WLLN for sample means, along with Proposition 3, we have $\lim_{T\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_T,\mathcal{C}_T)=1$. So the second term on the right side of (42) is asymptotically negligible, and upon sending $\delta\downarrow 0$ and optimizing with respect to $\widetilde{\theta}_1$, we have

$$\liminf_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_2(T)\geq(1-\gamma)T)}{\log(T)}\geq-\inf_{\widetilde{\theta}_1\in\Theta\,:\,\widetilde{\theta}_1<\theta_2}\frac{D(\widetilde{\theta}_1,\theta_1)}{D(\widetilde{\theta}_1,\theta_2)}. \tag{43}$$

The desired uniform convergence in (5) follows from the fact that for each T, $x\mapsto\log\mathbb{P}_{\nu\pi}(N_2(T)\geq x)/\log(x)$ is a monotone decreasing function for $x>1$.

We now establish (6). Using Lemma 1, the right side of (43) is equal to -1. Then using Assumption 1 and Markov's inequality, the case $x=(1-\gamma)T$ in (6) is established:

$$\lim_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_2(T)\geq(1-\gamma)T)}{\log((1-\gamma)T)}=-1. \tag{44}$$

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

$$\mathbb{P}_{\nu\pi}(N_2(T)\geq T^\gamma)\geq\mathbb{P}_{\nu\pi}(N_2(\lceil T^\gamma/(1-\gamma)\rceil)\geq T^\gamma). \tag{45}$$

Using (44), but with $\lceil T^\gamma/(1-\gamma)\rceil$ in the place of T, together with Assumption 1 and Markov's inequality,

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

Thus, using (45), Assumption 1 and Markov's inequality, the case $x=T^\gamma$ in (6) is established:

$$\lim_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_2(T)\geq T^\gamma)}{\log(T^\gamma)}=-1. \tag{46}$$

Since, for each T, $x\mapsto\log\mathbb{P}_{\nu\pi}(N_2(T)\geq x)/\log(x)$ is a monotone decreasing function for $x>1$, the desired uniform convergence in (6) for $x\in[T^\gamma,(1-\gamma)T]$ follows from the matching limits at the endpoints $x=(1-\gamma)T$ and $x=T^\gamma$, established in (44) and (46), respectively.

We now show (5) for sub-optimal arm $k\geq 3$. Consider an alternative environment $\widetilde{\nu}=(\widetilde{\theta}_1,\ldots,\widetilde{\theta}_{k-1},\theta_k,\ldots,\theta_K)$, where $\widetilde{\theta}_j<\theta_k$ for all $1\leq j\leq k-1$. The events $\mathcal{B}_T$ and $\mathcal{C}_T$ become:

$$\mathcal{B}_T=\left\{\left|\widehat{\mu}_j(T)-\mu(\widetilde{\theta}_j)\right|\leq\delta\ \ \forall\,1\leq j\leq k-1\right\}$$
$$\mathcal{C}_T=\left\{\frac{N_j(T)}{\log(T)}\leq\frac{1+\delta}{D(\widetilde{\theta}_j,\theta_k)}\ \ \forall\,1\leq j\leq k-1\right\}\cap\left\{\frac{N_j(T)}{\log(T)}\leq\frac{1+\delta}{D(\theta_j,\theta_k)}\ \ \forall\,j\geq k+1\right\}.$$

To obtain (5) for sub-optimal arm $k\geq 3$, we can then run through arguments analogous to those in (37)-(43). Here, the change of measure from ν to $\widetilde{\nu}$ involves the product of $k-1$ likelihood ratios corresponding to the arms $1,\ldots,k-1$. Each of the parameter values $\widetilde{\theta}_1,\ldots,\widetilde{\theta}_{k-1}$ can be optimized separately to yield the desired conclusion. $\square$

6.2. Proof of Theorem 2

Without loss of generality, we assume that the (one-dimensional) exponential family model $\mathcal{M}$ is parameterized by the distribution mean. We will use μ_j to denote the mean of arm j. Also, without loss of generality, suppose that $\mu_1 > \mu_2 > \cdots > \mu_K$ (i.e., $k(i) = i$ for all $1 \le i \le K$). Define for arm k the KL-UCB index at time t, given that arm k has been played n times:

$$U_k(n,t) = \sup\left\{\mu : D(\widehat{\mu}_k(\tau_k(n)), \mu) \le \frac{\log(t)}{n}\right\},$$

where $\tau_k(n)$ is time of the n-th play of arm k. Note that we use the choice $f(t) = \log(t)$ in KL-UCB, where the "exploration function" $f(t)$ is a design choice in Algorithm 2 of Cappé et al. (2013). In particular, Section 7 of Cappé et al. (2013) recommends using this choice of $f(t)$. However, our proof below can easily accommodate other choices such as $f(t) = \log(t) + 3\log(\log(t))$, which Theorem 1 of Cappé et al. (2013) assumes.

We first show (14) for the sub-optimal arm $k = 2$. 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:

$$\begin{aligned}
\mathbb{P}_{\nu\pi}\left(N_2(T) > x_T\right) &= \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (\tau_2(x_T), T] \;\text{ s.t. }\; U_1(N_1(t-1), t) < U_2(N_2(t-1), t)\Big) && (47)\\
&\le \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \;\text{ s.t. }\; U_1(N_1(t-1), x_T) < U_2(x_T, T)\Big)\\
&\le \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \;\text{ s.t. }\; U_1(N_1(t-1), x_T) \le \mu_2 + \delta\Big) && (48)\\
&\quad + \mathbb{P}_{\nu\pi}\Big(U_2(x_T, T) > \mu_2 + \delta\Big). && (49)
\end{aligned}$$

Note that (47) 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 (48), we have

$$\begin{aligned}
(48) &\le \sum_{i=1}^{\infty} \mathbb{P}_{\nu\pi}\Big(U_1(i, x_T) \le \mu_2 + \delta\Big) && (50)\\
&= \sum_{i=1}^{\infty} \mathbb{P}_{\nu\pi}\left(D(\widehat{\mu}_1(\tau_1(i)), \mu_2 + \delta) \ge \frac{\log(x_T)}{i},\;\; \widehat{\mu}_1(\tau_1(i)) \le \mu_2 + \delta\right)\\
&= \sum_{i=1}^{\infty} \mathbb{P}_{\nu\pi}\left(\frac{1}{i}\sum_{l=1}^{i} X_1(l) \le y_i^*\right), && (51)
\end{aligned}$$

where for each i, y_i^* is the unique solution to $D(y_i^*, \mu_2 + \delta) = \log(x_T)/i$ and $y_i^* < \mu_2 + \delta$. Then, continuing from (51), we have

$$\begin{aligned}
(51) &\le 2 \cdot \sum_{i=1}^{\infty} \exp\left(-i \cdot D(y_i^*, \mu_1)\right) && (52)\\
&\le 2 \cdot \sum_{i=1}^{\lfloor s_T \rfloor} \exp\left(-i \cdot D(y_i^*, \mu_2 + \delta) \cdot \inf_{z < \mu_2 + \delta} \frac{D(z, \mu_1)}{D(z, \mu_2 + \delta)}\right) && (53)
\end{aligned}$$

$$+2\cdot\sum_{i=\lfloor s_T\rfloor+1}^{\infty}\exp\left(-i\cdot\left(D(y_i^*,\mu_2+\delta)\cdot\inf_{z<\mu_2+\delta}\frac{D(z,\mu_1)}{D(z,\mu_2+\delta)}+\frac{D(\mu_2+\delta,\mu_1)}{2}\right)\right) \tag{54}$$

$$\leq x_T^{-\inf_{z<\mu_2+\delta}\frac{D(z,\mu_1)}{D(z,\mu_2+\delta)}}\cdot 2\cdot\left(s_T+\sum_{i=1}^{\infty}\exp\left(-i\cdot\frac{D(\mu_2+\delta,\mu_1)}{2}\right)\right), \tag{55}$$

where (as defined),

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

Note that in (52), we have applied a large deviations upper bound. In (53)-(54), after splitting into two summations, we have used the fact that $D(y_i^*,\mu_1)\geq D(\mu_2+\delta,\mu_1)$, and so for $i\geq s_T$, we have

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

which then yields

$$D(y_i^*,\mu_1)\geq\frac{\log(x_T)}{i}\cdot\inf_{z<\mu_2+\delta}\frac{D(z,\mu_1)}{D(z,\mu_2+\delta)}+\frac{D(\mu_2+\delta,\mu_1)}{2}.$$

For the term in (49), we have for sufficiently large T,

$$|U_2(x_T,T)-\widehat{\mu}_2(\tau_2(x_T))|<\frac{\delta}{2}.$$

So for sufficiently large T,

$$\begin{aligned}(49)&\leq\mathbb{P}_{\nu\pi}\left(\frac{1}{x_T}\sum_{l=1}^{x_T}X_2(l)>\mu_2+\frac{\delta}{2}\right)\\&\leq 2\cdot\exp\Big(-x_T\cdot\Lambda^*(\delta/2)\Big),\end{aligned} \tag{56}$$

where Λ^* is the large deviations rate function for P_2.

Putting together (48), (55) and (49), (56), we have

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

From the argument included separately in Appendix C,

$$\lim_{\delta\downarrow 0}\inf_{z<\mu_2+\delta}\frac{D(z,\mu_1)}{D(z,\mu_2+\delta)}=\inf_{z<\mu_2}\frac{D(z,\mu_1)}{D(z,\mu_2)}. \tag{57}$$

Using the lower bound for regret tail probabilities:

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

which was established in the proof of Theorem 1, 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 case $k=2$ is established

We now show (14) for sub-optimal arm $k \geq 3$. In parallel to (48) and (49), we have

$$\mathbb{P}_{\nu\pi}\left(N_k(T) > x_T\right) \leq \mathbb{P}_{\nu\pi}\Big(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \max_{1\leq j\leq k-1} U_j(N_j(t), x_T) \leq \mu_k + \delta\Big) \tag{58}$$

$$+\, \mathbb{P}_{\nu\pi}\Big(U_k(x_T, T) > \mu_k + \delta\Big). \tag{59}$$

We can bound (58) via:

$$(58) \leq \mathbb{P}_{\nu\pi}\Big(\forall\, 1 \leq j \leq k-1, \;\; \exists\, i(j) \in \mathbb{Z}_+ \;\text{ s.t. }\; U_j(i(j), x_T) \leq \mu_k + \delta\Big)$$

$$\leq \prod_{j=1}^{k-1} \sum_{i(j)=1}^{\infty} \mathbb{P}_{\nu\pi}\Big(U_j(i(j), x_T) \leq \mu_k + \delta\Big), \tag{60}$$

where (60) follows from the independence of the rewards from different arms. We can then upper bound each term in the product of (60) in the same way as (50). We can upper bound (59) in the same way as (49), and thus show that it is asymptotically negligible. Following the rest of the argument above for the case $k = 2$, we eventually obtain (due to the product structure in (60)):

$$\limsup_{T\to\infty} \frac{\log \mathbb{P}_{\nu\pi}\left(N_k(T) > x_T\right)}{\log(x_T)} \leq -\sum_{j=1}^{k-1} \inf_{z<\mu_k} \frac{D(z, \mu_j)}{D(z, \mu_k)}$$

The theorem conclusion is established by the matching lower bounds established in (5) of Theorem 1, 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). $\square$

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 and the resulting exponent of the regret distribution tail, established in (31) and described in Remark 3.

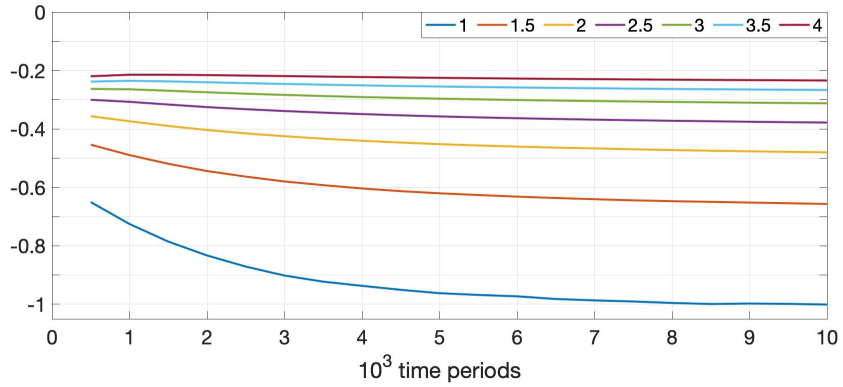

Figure 1 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) \geq 0.8T)/\log(T)$ vs T. Environment $\nu = (N(0.1, \sigma_0^2), N(0, \sigma_0^2))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards. The curves correspond to the cases $\sigma_0^2 = 1, 1.5, \ldots, 4$, as indicated by the legend. The curves asymptote to $-1/\sigma_0^2$ in each case, which agrees with Corollary 2 and (21).

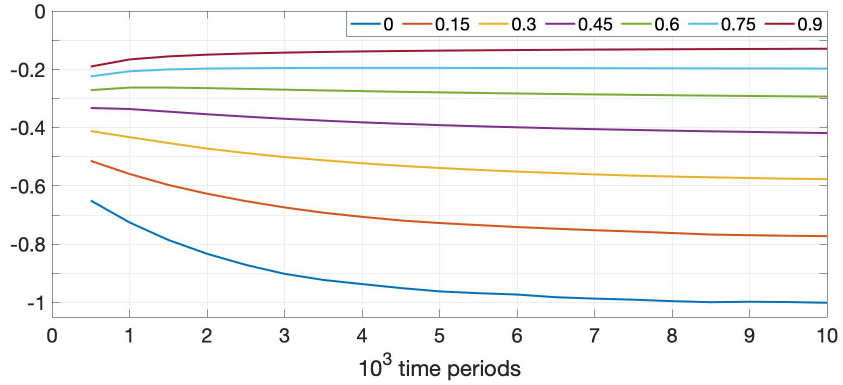

Figure 2 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) \geq 0.8T)/\log(T)$ vs T. Environment ν consists of two Gaussian AR(1) processes with common AR coefficient β_0, and equilibrium distributions $(N(0.1, 1), N(0, 1))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards. The curves correspond 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 (24).

In Figure 5, we demonstrate that the mis-specification of the serial dependence structure of the rewards, which is issue (with the same AR(1) setup) that is illustrated in Figure 2, can be overcome using the robustness modifications described in Section 5.2. As discussed in first paragraph of Remark 4, here we do not have upper bounds on regret tail probabilities (only lower bounds in (24)), and thus there are no provable robustness guarantees. However, we show in Figure 5 that the robustness modifications still work well empirically. The $\frac{1+\beta_0}{1-\beta_0}$ factor in Figure 5 is taken from the lower bound in (24), which we essentially confirm to be tight here.

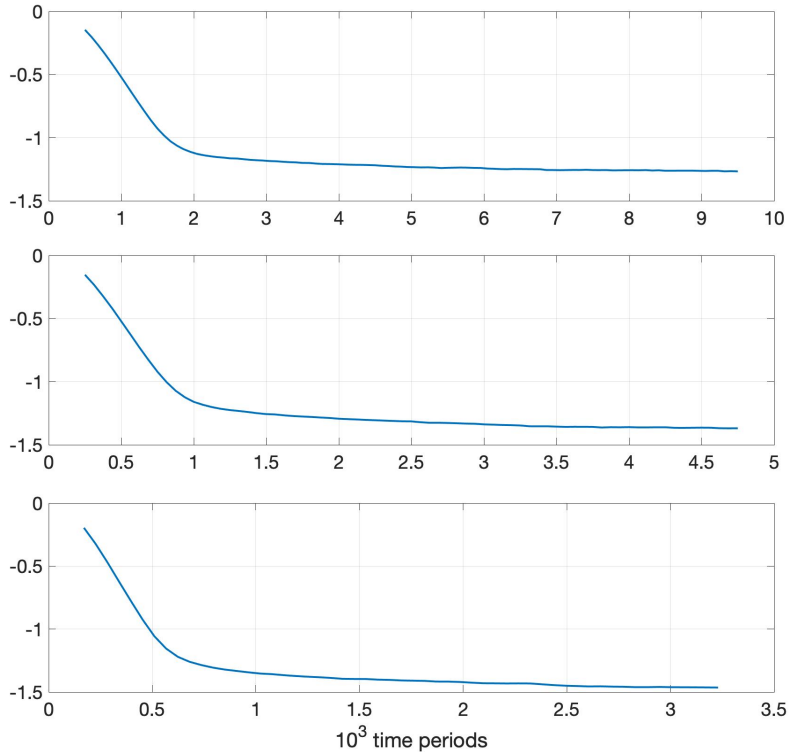

Figure 3 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$ (with time horizon T fixed). Environment $\nu = (\mathrm{Ber}(p), \mathrm{Ber}(0.4))$. Algorithm π is KL-UCB for iid Bernoulli rewards. Top: $p = 0.475$, $T = 10^4$; Middle: $p = 0.5$, $T = 5 \times 10^3$;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

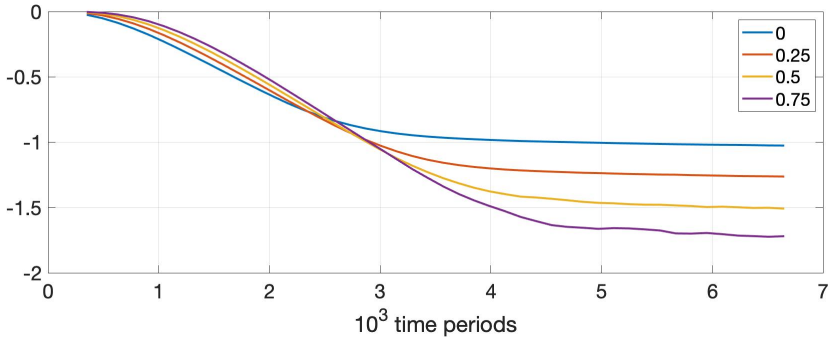

Figure 4 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$, with fixed time horizon $T = 7 \times 10^3$. Environment $\nu = (N(0.1,1), N(0,1))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards, but with the amount of exploration increased by $(1+b)$ times, to aim for a regret tail exponent of $-(1+b)$. The curves correspond to the cases $b = 0, 0.25, 0.5, 0.75$, as indicated by the legend.

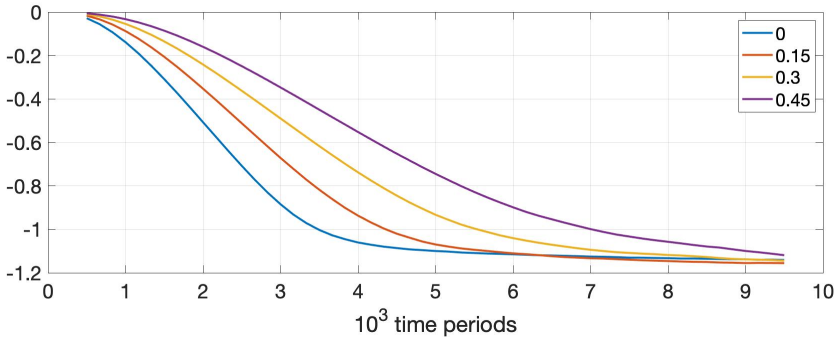

Figure 5 Plot of $\log \mathbb{P}_{\nu\pi}(N_2(T) > x)/\log(x)$ vs x for $x \in [0.05T, 0.95T]$, with fixed time horizon $T = 10^4$. Environment ν consists of two Gaussian AR(1) processes with common AR coefficient β_0, and equilibrium distributions $(N(0.1,1), N(0,1))$. Algorithm π is KL-UCB for iid unit-variance Gaussian rewards, but with the amount of exploration increased by $1.1 \cdot \frac{1+\beta_0}{1-\beta_0}$ times, to aim for a regret tail exponent of ≈ -1.1 for environment ν (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.

Appendix A: Proofs for Section 3.1

Proof of Lemma 1. We first show the forward direction, that Assumption 2 implies (7). From (2), we have

$$\frac{d}{dy} D(x,y) = A'(y) - A'(x) > 0$$

for all $y > x$, since A' is a strictly increasing function. Thus, for any θ_1, θ_2 fixed with $\theta_1 > \theta_2$,

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

There are two possible cases:

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

In the first case, note that for any fixed non-zero θ_0, we have by L'Hôpital's rule,

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

And if $\theta_0 = 0$, then of course the same limit result holds. So Assumption 2 implies that

$$\lim_{\theta \to -\infty} \frac{D(\theta, \theta_1)}{D(\theta, \theta_2)} = \lim_{\theta \to -\infty} \frac{\theta A'(\theta) - A(\theta)}{\theta A'(\theta) - A(\theta)} = 1. \tag{61}$$

In the second case, Assumption 2 directly implies (61). Thus, the forward direction is established.

Now we show the reverse direction, starting with (7). Suppose $\inf \Theta > -\infty$. Note that (7) implies that for any fixed $\theta_0 > \inf \Theta$,

$$\lim_{\theta \downarrow \inf \Theta} D(\theta, \theta_0) = \lim_{\theta \downarrow \inf \Theta} A(\theta_0) - A(\theta) - A'(\theta) \cdot (\theta_0 - \theta) = \infty.$$

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

$$\lim_{\theta\downarrow\inf\Theta} A(\theta)+\epsilon A'(\theta)=-\infty \tag{62}$$

for any $\epsilon>0$. Because $\inf\Theta>-\infty$, (62) implies that

$$\lim_{\theta\downarrow\inf\Theta} A'(\theta)=-\infty, \tag{63}$$

since A is strictly convex and A' is strictly increasing on Θ. Since

$$\lim_{\theta\downarrow\inf\Theta} A(\theta)>-\infty,$$

(62)-(63) imply that

$$\lim_{\theta\downarrow\inf\Theta}\left|\frac{A(\theta)}{A'(\theta)}\right|=0.$$

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

$$\lim_{\theta\downarrow\inf\Theta}\frac{D(\theta,\theta_1)}{D(\theta,\theta_2)}=\lim_{\theta\downarrow\inf\Theta}\frac{A'(\theta)\cdot(\theta_1-\theta)}{A'(\theta)\cdot(\theta_2-\theta)}=\frac{\theta_1-\inf\Theta}{\theta_2-\inf\Theta}>1,$$

which contradicts (7) if $\inf\Theta>-\infty$. Hence, it must be that $\inf\Theta=-\infty$.

Now suppose that

$$\lim_{\theta\to-\infty}\theta A'(\theta)-A(\theta)<\infty. \tag{64}$$

Again, consider two the possible cases:

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

In the first case, (7) cannot hold because (64) implies (for $\theta_1>\theta_2$):

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

In the second case, (7) cannot hold because (64) then implies that $\lim_{\theta\to-\infty}D(\theta,\theta')<\infty$ for any $\theta'\in\Theta$. So it must be that

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

We have thus established the reverse direction, that (7) implies Assumption 2. □

Proof of Proposition 1. Since $\inf\Theta = -\infty$ 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 $\mu(\theta) = A'(\theta)$ must also go to $-\infty$. By the definition of the convex conjugate A^*, we have for any $\theta \in \Theta$,

$$A^*(x) \geq \theta \cdot x - A(\theta),$$

which implies for $\theta < 0$ that

$$\lim_{x \to -\infty} A^*(x) = \infty.$$

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

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

So using $\lim_{\theta \to -\infty} A'(\theta) = -\infty$ yields the desired result. $\square$

Proof of Proposition 2. Let X be a random variable with distribution h, where h is from the model (1). 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:

$$\begin{aligned} A^*(l) &= \sup_{\theta \in \Theta} \big(\theta \cdot l - \log \mathbb{E}[\exp(\theta X)]\big) \\ &= -\log\left(\inf_{\theta \in \Theta} \mathbb{E}[\exp(\theta(X - l))]\right). \end{aligned} \tag{65}$$

Let us decompose via:

$$\mathbb{E}[\exp(\theta(X-l))] = \mathbb{E}[\exp(\theta(X-l)); X \geq l] + \mathbb{E}[\exp(\theta(X-l)); X < l]. \tag{66}$$

For $\theta < 0$, the first term on the right side of (66) can be bounded via:

$$0 \leq \mathbb{E}[\exp(\theta(X-l)); X \geq l] \leq \exp(|\theta(l-L)|) \cdot \mathbb{E}[\exp(\theta(X-L))], \tag{67}$$

while the second term can be bounded via:

$$0 \leq \mathbb{E}[\exp(\theta(X-l)); X < l] \leq \exp(|\theta(l-L)|) \cdot \mathbb{P}(X < l). \tag{68}$$

Now for any $\epsilon > 0$, set $\theta = -1/\epsilon$ and $l - L = \epsilon$, so that $\exp(|\theta(l-L)|) = \exp(1)$ in (67)-(68). 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 (67)-(68) can be made arbitrarily small by taking (the just defined) $\epsilon > 0$ to be sufficiently small. Therefore, we have shown that

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

which, by (65), translates into

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

Since

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

we have

$$\lim_{\theta \to -\infty} A^*(A'(\theta)) = \infty, \tag{69}$$

which is the equivalent representation for Assumption 2.

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 $\inf_{\theta \in \Theta} \mathbb{E}[\exp(\theta(X - l))]$ from (65) 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:

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

Therefore, we have shown that

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

which, by (65), translates into

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

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

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

unlike in the case of continuous distributions, where we ended up with (69), here we have

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

□

Appendix B: Proofs for Section 3.2

Proof of Proposition 4. Let $k \geq 2$. From the lower bounds in (43) in the proof of Theorem 1, there exists $a > 0$ such that for all $x \in [T^\gamma, (1-\gamma)T]$ and T sufficiently large,

$$\begin{aligned} T^{-a} &\leq \mathbb{P}_{\nu\pi}(N_k(T) \geq (1-\gamma)T) \\ &\leq \mathbb{P}_{\nu\pi}(N_k(T) \geq x). \end{aligned}$$

Thus,

$$\begin{aligned} 0 &\leq \mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_k(T) - \mu(\theta_k)| > \epsilon \mid N_k(T) \geq x\right) \\ &\leq \frac{\mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_k(T) - \mu(\theta_k)| > \epsilon,\ N_k(T) \geq T^\gamma\right)}{\mathbb{P}_{\nu\pi}\left(N_k(T) \geq (1-\gamma)T\right)} \\ &\leq T^a \cdot 2\exp\left(-T^\gamma \cdot (\Lambda^*(\epsilon) \wedge \Lambda^*(-\epsilon))\right), \end{aligned}$$

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, with Λ^* being the large deviations rate function. So $\mathbb{P}_{\nu\pi}\left(|\widehat{\mu}_k(T) - \mu(\theta_k)| > \epsilon \mid N_k(T) \geq x\right) \to 0$ uniformly for $x \in [T^\gamma, (1-\gamma)T]$ as $T \to \infty$, which yields the desired result. □

Appendix C: Proofs for Section 3.3

Verification of (57) in Proof of Theorem 2. From the identity (2), we have for any μ,

$$\frac{d}{dz} D(z, \mu) = -A''(z)(\mu - z).$$

So

$$\frac{d}{dz}\frac{D(z,\mu_1)}{D(z,\mu_2+\delta)} = \frac{A''(z)}{D(z,\mu_2+\delta)^2}\left(\underbrace{D(z,\mu_1)(\mu_2+\delta-z) - D(z,\mu_2+\delta)(\mu_1-z)}_{:=\xi(z)}\right).$$

Note that $\xi(\mu_2+\delta) = 0$ and $\xi'(z) = D(z,\mu_2+\delta) - D(z,\mu_1)$ for $z < \mu_2+\delta$. So $\xi'(z) < 0$, and thus $\xi(z) > 0$ for $z < \mu_2+\delta$. This, together with the fact that $A''(z) \geq 0$ for all z, we conclude that $z \mapsto D(z,\mu_1)/D(z,\mu_2+\delta)$ is monotone increasing for $z < \mu_2+\delta$. So for any $\delta > 0$,

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

Since $\delta \mapsto D(z,\mu_1)/D(z,\mu_2+\delta)$ is monotone decreasing, $\delta \mapsto \inf_{z<\mu_2} D(z,\mu_1)/D(z,\mu_2+\delta)$ is also monotone decreasing. Therefore,

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

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

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

□

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 $\nu = (P_1, \ldots, P_K) \in \mathcal{M}^K$ and any sub-optimal arm k, we have $N_k(T) \to \infty$ in $\mathbb{P}_{\nu\pi}$-probability as $T \to \infty$.*

Proof of Lemma 2. Suppose the conclusion is false for some environment $\widetilde{\nu} = (\widetilde{P}_1, P_2, \ldots, P_K) \in \mathcal{M}^K$. Without loss of generality, suppose arm 1 is sub-optimal in $\widetilde{\nu}$. So there exists $m > 0$, $\epsilon > 0$ and a deterministic sequence of times $T_n \uparrow \infty$ such that

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(N_1(T_n) \le m\right) \ge \epsilon. \tag{70}$$

Consider another environment $\nu = (P_1, P_2, \ldots, P_K) \in \mathcal{M}^K$, with $\mu(P_1) > \mu_*$. Define the events:

$$\mathcal{B}_n = \{N_1(T_n) \le m\}, \quad \mathcal{C}_n = \left\{ \frac{1}{N_1(T_n)} \sum_{t=1}^{N_1(T_n)} \log \frac{d\widetilde{P}_1}{dP_1}(X_1(t)) \le D(\widetilde{P}_1, P_1) + \delta \right\}.$$

Pick $\delta > 0$ large enough so that

$$\mathbb{P}_{\widetilde{\nu}\pi}\left(\forall j = 1, \ldots, m : \frac{1}{j} \sum_{t=1}^{j} \log \frac{d\widetilde{P}_1}{dP_1}(X_1(t)) \le D(\widetilde{P}_1, P_1) + \delta \right) \ge 1 - \frac{\epsilon}{2}. \tag{71}$$

Defining $\mathcal{A}_n$ as in the proof of Proposition 5, and following the same steps starting with (19),

$$\begin{aligned} \mathbb{P}_{\nu\pi}(\mathcal{A}_n) &\ge \mathbb{E}_{\widetilde{\nu}\pi}\left[\exp\left(-N_1(T_n) \frac{1}{N_1(T_n)} \sum_{t=1}^{N_1(T_n)} \log \frac{d\widetilde{P}_1}{dP_1}(X_1(t)) \right); \mathcal{B}_n, \mathcal{C}_n \right] \\ &\ge \exp\left(-\left(D(\widetilde{P}_1, P_1) + \delta \right) m \right) \cdot \mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_n, \mathcal{C}_n). \end{aligned} \tag{72}$$

From (70) and (71), we have $\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_n, \mathcal{C}_n) \ge \frac{\epsilon}{2}$ for all n. So from (72), for the particular environment ν, there exist $k' \in \{2, \ldots, K\}$ and $\gamma > 0$ such that $\mathbb{P}_{\nu\pi}(N_{k'}(T_n) \ge T_n/(2K)) \ge \gamma$ for all n. This violates the consistency of π, and so (70) cannot be true. □

Appendix E: Proofs for Section 4.1

Proof of Theorem 4. First, note the following relationship between the KL divergence and the convex conjugate: $D(P^z_{k(j)}, P_{k(j)}) = \Lambda^*_{k(j)}(z)$, where $\Lambda^*_{k(j)}$ is the convex conjugate of the CGF $\Lambda_{k(j)}$ of $P_{k(j)}$. The proof of the lower bound part of (20) then follows from Theorem 5 (which uses Proposition 6). To establish the upper bound part of (20), we can use the same proof of Theorem 2; see Section 6.2. In that proof, we replace D by d in the appropriate places. The only thing that needs to be checked is the analog of (57):

$$\lim_{\delta \downarrow 0} \inf_{z < \mu(P_{k(2)}) + \delta} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)}) + \delta)} = \inf_{z < \mu(P_{k(2)})} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)}))}. \tag{73}$$

(Below, we check (73) for $k(1)$ and $k(2)$. The same arguments apply for the other combinations of $k(i)$, $i \geq 3$ and $k(j)$, $1 \leq j \leq i-1$.) First, there exists a fixed $\Delta > 0$ (depending on $P_{k(1)}$ and $P_{k(2)}$) such that for all $\delta > 0$ sufficiently small, we have both:

$$\inf_{z<\mu(P_{k(2)})} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)}))} = \inf_{z<\mu(P_{k(2)})-\Delta} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)}))}, \tag{74}$$

$$\inf_{z<\mu(P_{k(2)})+\delta} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)})+\delta)} = \inf_{z<\mu(P_{k(2)})-\Delta} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)})+\delta)}. \tag{75}$$

Note that

$$z \mapsto \frac{d(z, \mu(P_{k(2)}))}{d(z, \mu(P_{k(2)})+\delta)}$$

is monotone decreasing for $z < \mu(P_{k(2)})$, which we deduce from the verification of (57) in proof of Theorem 2 in Appendix C. Also, we have

$$\lim_{\delta \downarrow 0} \sup_{z<\mu(P_{k(2)})-\Delta} \frac{d(z, \mu(P_{k(2)}))}{d(z, \mu(P_{k(2)})+\delta)} = \sup_{\delta>0} \sup_{z<\mu(P_{k(2)})-\Delta} \frac{d(z, \mu(P_{k(2)}))}{d(z, \mu(P_{k(2)})+\delta)} = 1.$$

$$\lim_{\delta \downarrow 0} \frac{d(\mu(P_{k(2)})-\Delta, \mu(P_{k(2)}))}{d(\mu(P_{k(2)})-\Delta, \mu(P_{k(2)})+\delta)} = 1.$$

Therefore, we have uniform convergence for $z < \mu(P_{k(2)}) - \Delta$:

$$\lim_{\delta \downarrow 0} \sup_{z<\mu(P_{k(2)})-\Delta} \left| \frac{d(z, \mu(P_{k(2)}))}{d(z, \mu(P_{k(2)})+\delta)} - 1 \right| = 0. \tag{76}$$

For any $\epsilon \in (0,1)$, using (76), we have for sufficiently small $\delta > 0$:

$$\begin{aligned}(1-\epsilon) \inf_{z<\mu(P_{k(2)})-\Delta} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)}))} &\leq \inf_{z<\mu(P_{k(2)})-\Delta} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)})+\delta)} \cdot \frac{d(z, \mu(P_{k(2)}))}{d(z, \mu(P_{k(2)}))} \\ &\leq (1+\epsilon) \inf_{z<\mu(P_{k(2)})-\Delta} \frac{D(P^z_{k(1)}, P_{k(1)})}{d(z, \mu(P_{k(2)}))}.\end{aligned}$$

Sending $\delta \downarrow 0$, followed by $\epsilon \downarrow 0$, and then using (74)-(75), we obtain (73). □

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., $k(i) = i$ for $i = 1, 2$). The proof of the lower bound part of (20) follows from Theorem 5 (which uses Proposition 6). The upper bound part:

$$\limsup_{T \to \infty} \frac{\log \mathbb{P}_{\nu\pi}(N_2(T) \geq \log^{1+\gamma}(T))}{\log(\log^{1+\gamma}(T))} \leq -\frac{\sigma^2}{\sigma_0^2}, \tag{77}$$

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 $x \mapsto \log \mathbb{P}_{\nu\pi}(N_2(T) \geq x)/\log(x)$ for fixed T and $x > 1$.

□

Verification of (77) 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:

$$\begin{aligned}
&\mathbb{P}_{\nu\pi}\left(N_2(T) > x_T\right) \\
&= \mathbb{P}_{\nu\pi}\left(\exists\, t \in (\tau_2(x_T), T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(t)}{N_1(t-1)}} < \widehat{\mu}_2(t-1) + \sqrt{\frac{2\sigma^2 \log(t)}{N_2(t-1)}}\right) \\
&\le \mathbb{P}_{\nu\pi}\left(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(x_T)}{N_1(t-1)}} < \widehat{\mu}_2(\tau_2(x_T)) + \sqrt{\frac{2\sigma^2 \log(T)}{x_T}}\right) \\
&\le \mathbb{P}_{\nu\pi}\left(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(x_T)}{N_1(t-1)}} \le \mu_2 + \frac{\Delta}{2}\right) \qquad (78) \\
&\quad + \mathbb{P}_{\nu\pi}\left(\widehat{\mu}_2(\tau_2(x_T)) + \sqrt{\frac{2\sigma^2 \log(T)}{x_T}} > \mu_2 + \frac{\Delta}{2}\right). \qquad (79)
\end{aligned}$$

For the term in (78), we have

$$\begin{aligned}
(78) &= \mathbb{P}_{\nu\pi}\left(\exists\, t \in (x_T, T] \;\text{ s.t. }\; \widehat{\mu}_1(t-1) + \sqrt{\frac{2\sigma^2 \log(x_T)}{N_1(t-1)}} \le \mu_1 - \frac{\Delta}{2}\right) \\
&\le \sum_{j=1}^{\infty} \mathbb{P}_{\nu\pi}\left(\frac{1}{j}\sum_{i=1}^{j} X_1(i) \le \mu_1 - \sqrt{\frac{2\sigma^2 \log(x_T)}{j}} - \frac{\Delta}{2}\right) \qquad (80) \\
&\le 2 \cdot \sum_{j=1}^{\infty} \exp\left(-\frac{j}{2\sigma_0^2}\left(\sqrt{\frac{2\sigma^2 \log(x_T)}{j}} + \frac{\Delta}{2}\right)^2\right) \qquad (81) \\
&= x_T^{-\sigma^2/\sigma_0^2} \cdot 2 \cdot \sum_{j=1}^{\infty} \exp\left(-\frac{\sqrt{j\sigma^2 \log(x_T)}\Delta}{\sqrt{2}\sigma_0^2} - \frac{j\Delta^2}{8\sigma_0^2}\right) \\
&\le x_T^{-\sigma^2/\sigma_0^2} \cdot 2 \cdot \sum_{j=1}^{\infty} \exp\left(-\frac{\sqrt{j}\sigma\Delta}{\sqrt{2}\sigma_0^2} - \frac{j\Delta^2}{8\sigma_0^2}\right) \qquad (\text{for } T \ge 16), \qquad (82)
\end{aligned}$$

where to obtain (80), we have used a union bound over all possible values of $N_1(t)$, $t \ge 1$, and to obtain (81), we have used a large deviations upper bound.

For the term in (79), we have for sufficiently large T,

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

So for sufficiently large T,

$$\begin{aligned}
(79) &\le \mathbb{P}_{\nu\pi}\left(\frac{1}{x_T}\sum_{i=1}^{x_T} X_2(i) > \mu_2 + \frac{\Delta}{4}\right) \\
&\le 2 \cdot \exp\left(-x_T \cdot \frac{\Delta^2}{32\sigma_0^2}\right), \qquad (83)
\end{aligned}$$

where to obtain (83), we have used a large deviations upper bound.

Putting together (78), (82) and (79), (83), 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 6. This proof is an extension and simplification of Propositions 7-8 of Cowan and Katehakis (2019).

Because Assumptions 3-4 hold, we have (22) by the Gärtner-Ellis Theorem for sample means and the Borel-Cantelli Lemma. Concerning (22), in the context of this proof, we will for simplicity denote the almost sure limit for arm i by μ_i. Recall that the UCB index for each arm i at time $t+1$ is defined via:

$$U_i(N_i(t), t+1) = \sup\left\{\mu : d(\widehat{\mu}_i(t), \mu) \leq \frac{\log(t+1)}{N_i(t)}\right\}, \tag{84}$$

and $\tau_i(m)$ denotes the time of the m-th play of arm i.

Without loss of generality, suppose that arm 1 is the unique optimal arm, i.e., $k(1) = 1$. We begin with the upper bound part of the proof. Let $\delta \in (0, (\mu_1 - \mu_k)/2)$. For each sub-optimal arm $k \geq 2$, we have

$$N_k(T) = 1 + \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = k,\, U_k(N_k(t), t+1) \geq \mu_1 - \delta,\, \widehat{\mu}_k(t) \leq \mu_k + \delta\right) \tag{85}$$

$$+ \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = k,\, U_k(N_k(t), t+1) \geq \mu_1 - \delta,\, \widehat{\mu}_k(t) > \mu_k + \delta\right) \tag{86}$$

$$+ \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = k,\, U_k(N_k(t), t+1) < \mu_1 - \delta\right), \tag{87}$$

where $\pi(t)$ is the arm sampled by the algorithm π at time t.

The first sum is upper bounded via:

$$(85) \leq \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = k,\, d(\mu_k + \delta, \mu_1 - \delta) \leq \frac{\log(t+1)}{N_k(t)}\right) \tag{88}$$

$$\leq \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1) = k,\, N_k(t) \leq \frac{\log(T)}{d(\mu_k + \delta, \mu_1 - \delta)}\right)$$

$$\leq \frac{\log(T)}{d(\mu_k + \delta, \mu_1 - \delta)} + 1. \tag{89}$$

The bound in (88) holds due to the events $U_k(N_k(t), t+1) \geq \mu_1 - \delta$ and $\widehat{\mu}_k(t) \leq \mu_k + \delta$ and the definition of the index in (84).

The second sum is upper bounded via:

$$(86) \leq \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1)=k,\, \widehat{\mu}_k(t) > \mu_k + \delta\right). \tag{90}$$

The sum in (90) can equal 1 for at most finitely many t by the SLLN for the sample mean and the fact that for each 1 in the sum, sub-optimal arm $k \geq 2$ is played an additional time and there is an additional sample that is averaged in the sample mean.

The third sum is upper bounded via:

$$\begin{aligned}(87) &\leq \sum_{t=K}^{T-1} \mathbb{I}\left(A(t+1)=k,\, U_1(N_1(t),t+1) \leq U_k(N_k(t),t+1) < \mu_1 - \delta\right) \\ &\leq \sum_{t=K}^{T-1} \mathbb{I}\left(U_1(N_1(t),t+1) < \mu_1 - \delta\right). \end{aligned} \tag{91}$$

The sum in (91) can equal 1 for at most finitely many t by the SLLN for the sample mean and the form of the index in (84) with $\log(t)$ increasing.

Putting together (89), (90) and (91) and sending $\delta \downarrow 0$, we have established that almost surely for each sub-optimal arm $k \geq 2$,

$$\limsup_{T\to\infty} \frac{N_k(T)}{\log(T)} \leq \frac{1}{d(\mu_k,\mu_1)}. \tag{92}$$

Therefore, for the optimal arm 1, almost surely,

$$\lim_{T\to\infty} \frac{N_1(T)}{T} = 1, \tag{93}$$

which then implies by the form of the index in (84) and the SLLN for the sample mean that almost surely,

$$\lim_{t\to\infty} U_1(N_1(t),t+1) = \mu_1.$$

This also implies that almost surely, all sub-optimal arms are played infinitely many times, due to the term $\log(t)$ growing without bound in the index (84).

We now develop the lower bound parts of the proof. For any m, $\tau_1(m)$ denotes the time of the m-th play of the optimal arm 1, and so for all sub-optimal arms $k \geq 2$,

$$U_1(N_1(\tau_1(m)-1),\tau_1(m)) > U_k(N_k(\tau_1(m)-1),\tau_1(m)). \tag{94}$$

We have for sufficiently large m, almost surely,

$$\max_{t\in[\tau_1(m),\tau_1(m+1)]} \frac{\log(t)}{N_k(t)} \leq \frac{\log(\tau_1(m+1))}{N_k(\tau_1(m)-1)}$$

$$
\begin{aligned}
&= \frac{\log(\tau_1(m+1))}{\log(\tau_1(m))} \frac{\log(\tau_1(m))}{N_k(\tau_1(m)-1)} \\
&\leq (1+\delta) \frac{\log(\tau_1(m))}{N_k(\tau_1(m)-1)} && (95) \\
&\leq (1+\delta) d(\mu_k - \delta, U_k(N_k(\tau_1(m)-1), \tau_1(m))) && (96) \\
&\leq (1+\delta) d(\mu_k - \delta, U_1(N_1(\tau_1(m)-1), \tau_1(m))) && (97) \\
&\leq (1+\delta) d(\mu_k - \delta, \mu_1 + \delta). && (98)
\end{aligned}
$$

Note that (95) is due to (93), (96) is due to the SLLN for the sample mean of the sub-optimal arm $k \geq 2$ and the form of the index in (84), (97) is due to (94), and (98) is due to the SLLN for the sample mean of the optimal arm 1. Therefore, sending $m \to \infty$ and $\delta \downarrow 0$, almost surely,

$$
\liminf_{T\to\infty} \frac{N_k(T)}{\log(T)} \geq \frac{1}{d(\mu_k, \mu_1)},
$$

which together with (92), completes the proof. $\square$

Proof of Theorem 5. Without loss of generality, suppose that the reward distributions for arms $1, \ldots, K$ within the environment ν satisfy $\Lambda_1'(0) > \Lambda_2'(0) > \cdots > \lambda_K'(0)$ (i.e., $k(i) = i$ for all $1 \leq i \leq K$). Consider any sub-optimal arm $k \geq 2$. Let $\widetilde{\nu}$ be an alternative environment where the reward distribution structure remains the same for (sub-optimal) arms $k, k+1, \ldots, K$, but the distribution of $\sum_{i=1}^n X_j(i)$ for arm j with $1 \leq j \leq k-1$ is

$$
F_j^n(dx; \lambda_j) = \exp\left(\lambda_j \cdot x - n\Lambda_j^n(\lambda_j)\right) F_j^n(dx),
$$

where $F_j^n(dx)$ is the original distribution for $\sum_{i=1}^n X_j(i)$ in the environment ν. Moreover, for $1 \leq j \leq k-1$, we let $\lambda_j < 0$ such that $\Lambda_k'(0) > \Lambda_j'(\lambda_j)$. So, in the context of (22), arm k yields larger almost sure long-run average rewards than arm $1 \leq j \leq k-1$ in the environment $\widetilde{\nu}$.

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

$$
\begin{aligned}
\mathcal{B}_T &= \left\{ \left| \widehat{\mu}_j(T) - \Lambda_j'(\lambda_j)) \right| \leq \delta \ \ \forall\, 1 \leq j \leq k-1 \right\} \\
\mathcal{C}_T &= \left\{ \frac{1-\delta}{d(\Lambda_j'(\lambda_j), \Lambda_k'(0))} \leq \frac{N_j(T)}{\log(T)} \leq \frac{1+\delta}{d(\Lambda_j'(\lambda_j), \Lambda_k'(0))} \ \ \forall\, 1 \leq j \leq k-1 \right\} \\
&\quad \cap \left\{ \frac{1-\delta}{d(\Lambda_j'(0), \Lambda_k'(0))} \leq \frac{N_j(T)}{\log(T)} \leq \frac{1+\delta}{d(\Lambda_j'(0), \Lambda_k'(0))} \ \ \forall\, j \geq k+1 \right\}.
\end{aligned}
$$

Following steps analogous to (37)-(41) in the proof of Theorem 1, we then have for sufficiently large T,

$$
\begin{aligned}
&\mathbb{P}_{\nu\pi}(N_k(T) \geq (1-\gamma)T) \geq \mathbb{P}_{\nu\pi}(\mathcal{B}_T, \mathcal{C}_T) \\
&= \mathbb{E}_{\widetilde{\nu}\pi} \left[\exp\left(-\sum_{j=1}^{k-1} \left(\lambda_j \cdot \sum_{t=1}^{N_j(T)} X_j(t) + N_j(T) \cdot \Lambda_j^{N_j(T)}(\lambda_j) \right) \right); \mathcal{B}_T, \mathcal{C}_T \right]
\end{aligned}
$$

$$\geq \mathbb{E}_{\widetilde{\nu}\pi}\left[\exp\left(\sum_{j=1}^{k-1}\Big(-\lambda_j\cdot\big(\Lambda_j'(\lambda_j)-\delta\big)+\Lambda_j(\lambda_j)-\delta\Big)N_j(T)\right);\mathcal{B}_T,\mathcal{C}_T\right] \tag{99}$$

$$= \mathbb{E}_{\widetilde{\nu}\pi}\left[\exp\left(-\sum_{j=1}^{k-1}\Big(\Lambda_j^*(\Lambda_j'(\lambda_j))+\delta(1-\lambda_j)\Big)N_j(T)\right);\mathcal{B}_T,\mathcal{C}_T\right] \tag{100}$$

$$\geq \exp\left(-\sum_{j=1}^{k-1}\big(\Lambda_j^*(\Lambda_j'(\lambda_j))+\delta(1-\lambda_j)\big)\frac{1\pm\delta}{d(\Lambda_j'(\lambda_j),\Lambda_k'(0))}\log(T)\right)\cdot\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_T,\mathcal{C}_T). \tag{101}$$

Note that we have used the event $\mathcal{B}_T$ in (99) along with Assumption 3 and event $\mathcal{C}_T$. In (100), we have used the definition of the convex conjugate Λ_j^*. And we have used the event $\mathcal{C}_T$ in (101), where the $\pm\delta$ notation indicates either $+\delta$ or $-\delta$, whichever allows for the lower bound. We also note that $\lim_{T\to\infty}\mathbb{P}_{\widetilde{\nu}\pi}(\mathcal{B}_T,\mathcal{C}_T)=1$ by (22) (which also holds for arguments $\lambda_j\neq 0$) and Proposition 6. Then, taking logs, sending $T\to\infty$ and sending $\delta\downarrow 0$, we have

$$\liminf_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_k(T)\geq(1-\gamma)T)}{\log(T)}\geq-\sum_{j=1}^{k-1}\frac{\Lambda_j^*(\Lambda_j'(\lambda_j))}{d(\Lambda_j'(\lambda_j),\Lambda_k'(0))}.$$

For each $1\leq j\leq k-1$, this holds for any $\lambda_j<0$ such that $\Lambda_j'(\lambda_j)<\Lambda_k'(0)$, and thus,

$$\liminf_{T\to\infty}\frac{\log\mathbb{P}_{\nu\pi}(N_k(T)\geq(1-\gamma)T)}{\log(T)}\geq-\sum_{j=1}^{k-1}\inf_{z<\Lambda_k'(0)}\frac{\Lambda_j^*(z)}{d(z,\Lambda_k'(0))}.$$

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

References

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* .

Bubeck S, Cesa-Bianchi N (2012) Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* 5(1):1–122.

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, 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 A O annd Garivier (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.

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

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).

Lehmann E, Casella G (1998) *Theory of Point Estimation* (Springer).

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

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* .

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* .

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* .