Proceedings of Machine Learning Research vol 134:1–37, 2021 34th Annual Conference on Learning Theory

Regret Minimization in Heavy-Tailed Bandits

Shubhada Agrawal SHUBHADAIITD@GMAIL.COM
TIFR, Mumbai

Sandeep Juneja JUNEJA@TIFR.RES.IN
TIFR, Mumbai

Wouter M. Koolen WMKOOLEN@CWI.NL
Centrum Wiskunde & Informatica

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We revisit the classic regret-minimization problem in the stochastic multi-armed bandit setting when the arm-distributions are allowed to be heavy-tailed. Regret minimization has been well studied in simpler settings of either bounded support reward distributions or distributions that belong to a single parameter exponential family. We work under the much weaker assumption that the moments of order $(1+\epsilon)$ are uniformly bounded by a known constant B, for some given $\epsilon > 0$. We propose an optimal algorithm that matches the lower bound exactly in the first-order term. We also give a finite-time bound on its regret. We show that our index concentrates faster than the well-known truncated or trimmed empirical mean estimators for the mean of heavy-tailed distributions. Computing our index can be computationally demanding. To address this, we develop a batch-based algorithm that is optimal up to a multiplicative constant depending on the batch size. We hence provide a controlled trade-off between statistical optimality and computational cost.

Keywords: Multi-armed bandits, heavy-tailed distributions, confidence intervals, regret-minimization

1. Introduction

In this paper, we consider the problem of sequential allocation of resources in an uncertain environment. The player is presented with K *arms*, which correspond to K unknown probability distributions. When the player selects an arm, she observes a sample generated independently from the corresponding underlying distribution, called *reward*. The player's aim is to maximize the average cumulative reward, which is equivalent to minimizing the expected regret, defined to be the shortfall between the cumulative expected reward of the player and the expected reward collected by the policy playing the arm with the maximum mean in all the rounds.

Regret minimisation in the stochastic multi-armed bandit model was first studied by Thompson (1933) in the context of designing efficient clinical trials, and by Robbins (1952). Lai and Robbins (1985) proposed a lower bound on the expected regret for parametric distributions and gave a framework for optimal strategies in this setting. This lower bound was later generalized by Burnetas and Katehakis (1996) (see Lattimore and Szepesvári (2020, Chapter 16) for a proof of the lower bound). Since then this problem has been well studied in the literature (see, e.g., Auer et al. (2002); Agrawal (1995); Honda and Takemura (2010, 2011); Agrawal and Goyal (2012); Garivier and Cappé (2011); Kaufmann et al. (2012); Cappé et al. (2013b); Bubeck et al. (2013); Honda and Takemura (2015)). We refer the reader to Bubeck et al. (2012) for a survey on the extensive literature on this problem and its variations.

Apart from clinical trials, this regret-minimization framework finds applications in many other settings including in advertisement selection, recommendation systems, internet routing and congestion control, etc. Despite the vast body of literature, the setting where the underlying distributions are heavy-tailed (disributions for which the moment-generating function is not defined for any $\theta > 0$) has been largely unaddressed. In most of the previous work, authors restrict the arm distributions to have bounded support or belong to a restrictive single parameter exponential family (SPEF) of distributions. In many bandit application domains such as in financial markets, or in congestion control over networks, it is rarely the case that the arm distributions are either parametric or bounded. In particular, it is well known that the stock returns in developed economies follow heavy-tailed distributions that typically have finite moments of order at most 4. Higher moments are not guaranteed to exist. Furthermore, daily exchange rates and income and wealth distributions may have heavier tails with finite moments of order less than 2 (see, e.g., Nicolau and Rodrigues (2019)). Thus, it is important to understand and develop a general theory and efficient (both, computationally and statistically) algorithms that have wider applicability.

Recently, there has been some interest beyond SPEFs. Bubeck et al. (2012) consider the non-parametric class of "sub-ψ" distributions, where the convex function ψ bounds the log-moment generating function of the arm-distributions. The ψ-UCB algorithm proposed by the authors is order-optimal. Bubeck et al. (2013) propose algorithms which are optimal up to constants for heavy-tailed distributions. They show that by using more robust estimators for the mean, as compared to the empirical average, one can achieve sub-linear expected regret. The setting considered by the authors is closest to ours. We compare the performance of our algorithm to that proposed by Bubeck et al. (2013). Vakili et al. (2013) also consider bandits with heavy-tailed distributions. They propose a strategy which is based on dividing the time into interleaving sequences for exploration and exploitation. Lattimore (2017) considers distributions with a known uniform bound on the kurtosis. Cowan et al. (2018) and Cowan and Katehakis (2015) consider Gaussian bandits with unknown mean and variance, and uniform bandits with unknown support, respectively. However, none of their algorithms exactly match the lower bound on the expected regret to the first order.

As in Agrawal et al. (2020a), it can be shown that if no restrictions are imposed on the class of arm-distributions, then the lower bound on the expected regret is unbounded. To make the problem learnable, we allow for distributions that have their $(1+\epsilon)^{th}$-moment uniformly bounded by a constant, B, for $\epsilon > 0$. However, the existence of any higher moments is not guaranteed. In particular, we focus on the class

$$\mathcal{L}_B \triangleq \left\{ \eta \in \mathcal{P}(\Re) : \mathbb{E}_\eta |X|^{1+\epsilon} \leq B \right\}, \tag{1}$$

where $\mathcal{P}(\Re)$ denotes the collection of probability measures on $\Re$, $B > 0$ and $\epsilon > 0$ are known constants, and $\mathbb{E}_\eta |X|^{1+\epsilon} := \int |y|^{1+\epsilon} d\eta(y)$ denotes the $(1+\epsilon)^{th}$ moment of η. This is a standard assumption in literature on heavy-tailed distributions (see, e.g., Bubeck et al. (2013); L.A. et al. (2020)). Also see Agrawal et al. (2020b) for a discussion on methods for estimating ϵ and B in specific settings. Under this mild assumption on the arm-distributions, we develop an algorithm that suffers regret which asymptotically matches the lower bound exactly, up to the first order term. We also give a finite time analysis for its regret. We look at the computational complexity of the algorithm and demonstrate a trade-off in the expected regret and the computational cost suffered by the proposed algorithm.

As is common in the stochastic-bandit literature, the performance guarantees of the algorithm involve proving convergence results, which are typically a consequence of Chernoff-Hoeffding like

inequalities. However, direct application of Hoeffding's type results is not valid in our setting. We develop non-asymptotic concentration inequalities for functionals of probability measures that appear in the lower bound, which may also be of independent interest. Our approach for constructing an index for each arm can be used to get tight anytime-valid confidence intervals for means of heavy-tailed distributions. We show that these confidence intervals are at least as tight as those for popular robust estimators such as truncated/trimmed empirical means in specific settings. See Lugosi and Mendelson (2019) for popular estimators for a distribution's mean in the heavy-tailed setting. We also demonstrate numerically that our algorithm suffers significantly less regret compared to the Robust-UCB algorithm of Bubeck et al. (2013), which derives its index from the aforementioned estimators for mean.

Computing the index can be very costly in this generality. To address this, we propose a batched version of the algorithm that computes the index for each arm only at the beginning of each batch, and allocates all the samples within that batch to the arm with maximum value of the computed index. We show that with carefully chosen batch sizes, this batched-algorithm suffers regret that is off by only a constant multiplicative factor, while significantly improving the computational cost.

Since we allow for heavy-tailed distributions, we can no longer identify the distributions with one parameter, as is the case in the most widely used setting of SPEF. We work in the space of probability measures, where we use the Lévy metric (or equivalently the topology of weak-convergence) to define the notion of convergence.

We also establish the conjectured optimality of the Empirical KL-UCB algorithm of Cappé et al. (2013b) and give the first optimal finite-time regret bounds for bounded-support arm distributions. In Garivier and Cappé (2011), the authors gave a finite-time bound for a modification of the algorithm, and established its optimality only in the special case of Bernoulli arms. Maillard et al. (2011) independently gave a tight finite time analysis for the Bernoulli case. For general bounded-support distributions, Honda and Takemura (2010) proposed an asymptotically optimal algorithm, but did not give finite-time bounds on the regret.

In a nutshell, we propose the first asymptotically optimal algorithm for the heavy-tailed setting that matches the instance-dependent lower bound exactly up to the first order term. In this generality, the computational cost incurred by the algorithm can be significant. We propose a modification of the optimal algorithm that matches the lower bound up to constants, but requires significantly less computational effort. Our index suggests tight anytime-valid confidence intervals for the mean of heavy-tailed distributions, which are superior to those for the well-known truncation or trimming based estimators in specific settings. Moreover, the finite time analysis presented can be used to establish finite time guarantees for some of the existing algorithms.

Roadmap: In Section 2 we describe the setup and discuss the lower bound for the regret-minimization problem. Our proposed algorithm is presented in Section 3. Section 3.1 contains our main results for the proposed algorithm, including the finite time guarantee and its asymptotic optimality. A modification of the original algorithm that is practically tuned but at the cost of optimality up to constants, is also presented in this section. The regret analysis of the algorithm and the proof ideas for its theoretical guarantee are presented in Section 3.3. Superiority of our algorithm over that of Bubeck et al. (2013) is established in Section 3.4. In this section we also show exactly how our anytime-valid confidence intervals dominate those for popular mean-estimators for heavy-tailed distributions. We discuss the trade-off in the computational cost and statistical optimality of the algorithm in Section 3.5. We present the results of our numerical experiments in Section 4, and conclude in Section 5.

2. Background and the lower bound

Let $\mathcal{P}(\Re)$ denote the collection of all probability measures on $\Re$. For $\eta \in \mathcal{P}(\Re)$, let $m(\eta) = \int_{\Re} x d\eta$ denote the mean of distribution η. Furthermore, let $\mathrm{KL}(\eta, \kappa) = \int \log\left(\frac{d\eta}{d\kappa}\right) d\eta(x)$ denote the Kullbeck-Leibler divergence from probability distribution η to κ. Let $\mathcal{L} \subseteq \mathcal{P}(\Re)$ be any collection of probability measures, and let $\mathcal{L}^K$ denote the collection of vectors of K distributions, each from $\mathcal{L}$.

Given $\mu \in \mathcal{L}^K$ such that $\mu = \{\mu_1, \mu_2, \ldots, \mu_K\}$, we assume for convenience that arm 1 has the maximum mean. Let the arm selected by the algorithm at time n be denoted by A_n and let the observed random sample at that time be denoted by X_n. Then X_n is an independent sample distributed according to μ_{A_n}. Define the expected regret till time T as follows:

$$\mathbb{E}(R_T) \triangleq \sum_{n=1}^{T} m(\mu_1) - \mathbb{E}(m(\mu_{A_n})) = \sum_{a=1}^{K} \mathbb{E}(N_a(T))\, \Delta_a, \tag{2}$$

where $N_a(T)$ denotes the number of times arm a has been pulled in T trials, and $\Delta_a := m(\mu_1) - m(\mu_a)$ is the sub-optimality gap of arm a. For distribution $\eta \in \mathcal{P}(\Re)$ and candidate mean $x \in \Re$, we define

$$\mathrm{KL}^{\mathcal{L}}_{\inf}(\eta, x) := \inf\{\mathrm{KL}(\eta, \kappa) : \kappa \in \mathcal{L} \text{ and } m(\kappa) \geq x\}. \tag{3}$$

Then, Burnetas and Katehakis (1996) show that any reasonable strategy acting on a bandit problem $\mu \in \mathcal{L}^K$, for any collection $\mathcal{L}$, suffers expected regret satisfying:

$$\liminf_{T\to\infty} \frac{\mathbb{E}(R_T)}{\log(T)} \geq \sum_{a:\, m(\mu_a)<m(\mu_1)} \frac{\Delta_a}{\mathrm{KL}^{\mathcal{L}}_{\inf}(\mu_a, m(\mu_1))}, \tag{4}$$

where R_T denotes the total regret suffered by the algorithm till time T. The proof of lower bound (4) relies on change of measure arguments (see Lattimore and Szepesvári (2020)). Agrawal et al. (2020a, Lemma 1) show that it is necessary to impose certain restrictions on the class $\mathcal{L}$ under consideration, otherwise $\mathrm{KL}^{\mathcal{L}}_{\inf}(\cdot,\cdot) = 0$ leading to unbounded expected regret. To this end, we restrict $\mathcal{L}$ to the class $\mathcal{L}_B$ defined in (1), i.e., to the collection of all distributions satisfying $\mathbb{E}(|X|^{1+\epsilon}) \leq B$, for fixed positive constants ϵ and B. Notice that in order to bound the expected regret in (2), it is sufficient to bound the expected number of pulls of each sub-optimal arm $a \neq 1$.

Building upon the algorithms with KL-based confidence intervals in Maillard et al. (2011), Garivier and Cappé (2011), and Cappé et al. (2013b), where the authors propose optimal algorithms for much simpler settings of either Bernoulli or SPEF arms, we develop an algorithm that matches the lower bound in (4), in a much more general setting, allowing for heavy-tailed distributions.

As mentioned in the introduction, we study the convergence of sequences of probability measures in the Lévy metric, which is reviewed in Appendix A. The analysis of the algorithm uses the continuity and convexity properties of $\mathrm{KL}_{\inf}$, which are also proven in Appendix A (Lemma 10).

3. The algorithm

Our algorithm follows the UCB template with two variations. First, our arm indices are constructed from deviation inequalities for $\mathrm{KL}^{\mathcal{L}}_{\inf}$, which allows us to show that our algorithm matches the (asymptotic) instance-optimal regret lower bound for the heavy-tailed setting. Second, our algorithm processes the samples in batches. Our geometric batching allows us to reduce the worst-case run-time

from essentially $O(T^2)$ for the sample-at-a-time case to $O(T \log\log(T))$, at the cost of a mild factor in our regret.

Algorithm 1 formally describes our algorithm, $\mathrm{KL}_{\inf}$-UCB. This is an index-based algorithm that proceeds in batches. After each batch, it computes an index for each arm and allocates the next batch of samples to the arm with the maximum index (breaking ties arbitrarily, if any). Henceforth, unless specified, we set $\mathcal{L} = \mathcal{L}_B$, as defined in (1). Furthermore, when $\mathcal{L}$ is clear from the context, for ease of notation, we denote the functional $\mathrm{KL}_{\inf}^{\mathcal{L}}$, defined in (3), by $\mathrm{KL}_{\inf}$.

Let $\mu \in \mathcal{L}^K$ be the given bandit instance, $U_a(n)$ denote the value of the index corresponding to arm a at time n, A_n denote the arm selected at time n, and let $[K]$ denote the set $\{1, \ldots, K\}$. For simplicity of notation, in our algorithm and analysis, we denote by $N_a(n)$ the number of times arm a has been sampled in $n-1$ rounds. Moreover, let $\hat{\mu}_a(n)$ denote the empirical distribution corresponding to $N_a(n)$ samples from arm a, and $\tilde{\eta} \geq 0$ be a multiplicative factor that will be used to determine the batch sizes. The algorithm takes as inputs K, $\tilde{\eta}$, B, ϵ, and a threshold function, $g_a(.)$ corresponding to each arm, which will be used for computing the index for that arm. Our index

Input : K; description of $\mathcal{L}$, i.e., B and ϵ; $\tilde{\eta}$; threshold functions for each arm, i.e., $g_a(\cdot)$.

Initialization: Allocate 1 sample to each of the K arms.
Set $n \longleftarrow K+1$, $j \leftarrow K+1$.
Store empirical distributions, $\hat{\mu}_a(n)$, and update $N_a(n)$ for all arms $a \in [K]$.
while *True* **do**
  Compute index $U_a(n) = \sup\{x \in \Re : N_a(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_a(n), x) \leq g_a(n)\}$ for each arm.
  Compute best arm $A_n = \mathrm{argmax}_{a\in[K]}\, U_a(n)$ and batch size $B_j = \max\{1, \lceil \tilde{\eta} N_{A_n}(n) \rceil\}$.
  Sample arm A_n for B_j many trials and set $n \leftarrow n + B_j$, and $j \leftarrow j+1$.
  Update $\hat{\mu}_a(n)$ and $N_a(n)$ for each arm.
end

Algorithm 1: $\mathrm{KL}_{\inf}$ -UCB$(K, B, \epsilon, \tilde{\eta}, \{g_a(\cdot)\}_{a=1}^K)$.

$U_a(n)$ for arm a at time n is based on the functional $\mathrm{KL}_{\inf}$. It approximately corresponds to the inverse of $N_a(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_a(n), .)$, evaluated at the threshold, $g_a(n)$ and can be re-expressed as

$$U_a(n) = \max\{\mathbb{E}_\eta(X) :\ \eta \in \mathcal{L},\ N_a(n)\,\mathrm{KL}(\hat{\mu}_a(n), \eta) \leq g_a(n)\}\,. \tag{5}$$

It is the maximum mean among distributions in $\mathcal{L}$ that are close to the empirical distribution in KL divergence. This formulation of the index will be useful in comparing our confidence widths to those of the truncated empirical mean estimator (see Section 3.4). Before looking at the computational cost incurred by the algorithm, we look at its theoretical guarantees.

3.1. Main results

Let B_j be the random variable denoting the size of the j^{th} batch. Theorem 1 below gives a finite-time bound on the number of pulls of a sub-optimal arm by the proposed-algorithm for appropriately chosen threshold functions $g_a(\cdot)$. Corollary 2 shows that the $\mathrm{KL}_{\inf}$-UCB algorithm is asymptotically optimal up to a multiplicative factor of $(1+\tilde{\eta})$, and matches the lower bound when the batch size is 1, i.e., $\tilde{\eta} = 0$. However, from our discussion in Section 3.5, for $\tilde{\eta} = 0$, the algorithm has a quadratic computational cost. On the other hand, for $\tilde{\eta} > 0$, the computational cost reduces to being almost linear in the number of samples, at the cost of matching the lower bound upto a constant factor of

$1+\tilde{\eta}$. Thus, there is a trade off between the cost of computation and the expected regret suffered by the algorithm in the long run. In Section 4, numerical experiments demonstrate that even with the sub-optimality factor of $(1+\tilde{\eta})$, our algorithm suffers significantly less regret compared to the Robust-UCB algorithm with the truncated empirical mean as its estimator for the true mean, proposed in Bubeck et al. (2013). Our main regret bound is the following:

Theorem 1 ($\mathrm{KL}_{\inf}$**-UCB)** *Let $T > K \geq 2$, $\mu \in \mathcal{L}^K$, $\tilde{\eta} \geq 0$, and $g_a(t) = \log(t) + 2\log\log(t) + 2\log(1+N_a(t)) + 1$. For each sub-optimal arm a, $\mathrm{KL}_{\inf}$-UCB$(K, B, \epsilon, \tilde{\eta}, g_a(\cdot))$ has $\mathbb{E}(N_a(T))$ at most*

$$(1+\tilde{\eta})\left(\frac{\log T}{\mathrm{KL}_{\inf}(\mu_a, m(\mu_1))} + \frac{3(\log T)^{2/3}(c'_\mu)^{1/3}}{2(\mathrm{KL}_{\inf}(\mu_a, m(\mu_1)))^{4/3}} + O((\log T)^{1/3}) + O(\log\log(T))\right).$$

In Theorem 1 above, $c'_\mu > 0$ is a bandit instance-dependent constant. The exact $O((\log(T))^{1/3})$ and $O(\log\log T)$ terms are given in (11) below. Theorem 1 implies logarithmic regret for $\mathrm{KL}_{\inf}$-UCB, and Corollary 2 shows its asymptotic instance-optimality.

Corollary 2 *For $\mu \in \mathcal{L}^K$, $\tilde{\eta} \geq 0$, and $g_a(t) = \log(t) + 2\log\log(t) + 2\log(1+N_a(t)) + 1$, $\mathrm{KL}_{\inf}$-UCB, with inputs $(K, B, \epsilon, \tilde{\eta}, g_a(.))$ is asymptotically optimal up to a factor of $(1+\tilde{\eta})$, i.e.,*

$$\limsup_{T\longrightarrow\infty} \frac{\mathbb{E}(N_a(T))}{\log(T)} \leq \frac{1+\tilde{\eta}}{\mathrm{KL}_{\inf}(\mu_a, m(\mu_1))}.$$

Numerically we observe that the threhold $g_a(t)$ used by $\mathrm{KL}_{\inf}$-UCB is conservative for finite horizons. A version with much aggressive threshold of $\log t$ for all arms, performs much better (Figure 1). To address this, we propose a closely related algorithm, $\mathrm{KL}_{\inf}$-UCB2, which has threshold that is much smaller than $g_a(t)$, initially. This algorithm differs from $\mathrm{KL}_{\inf}$-UCB in that it solves the regret-minimization problem with respect to a slightly perturbed (larger) class while allowing for a more practically-relevant threshold. This results in smaller regret initially for sometime, at the cost of being asymptotically sub-optimal. Formally, let $\epsilon_1 > 0$, let $\tilde{B} = B + \epsilon_1$ and define $\delta_t = \log(1 + (\log\log(t))^{-1})$. For $\mu \in \mathcal{L}_B$, $\mathrm{KL}_{\inf}$-UCB2 is precisely $\mathrm{KL}_{\inf}$-UCB$(K, \tilde{B}, \epsilon, \tilde{\eta}, (1+\delta_t)^2\log(t))$. Notice that the threshold used here is much smaller than $g_a(t)$ for practically relavant horizons. For $\eta \in \mathcal{P}(\Re)$, and $x \in \Re$, let $\mathrm{KL}^{\epsilon_1}_{\inf}(\eta, x)$ denote $\mathrm{KL}^{\mathcal{L}_{\tilde{B}}}_{\inf}(\eta, x)$.

Theorem 3 ($\mathrm{KL}_{\inf}$**-UCB2)** *For $\mu \in \mathcal{L}^K$, $\tilde{\eta} \geq 0$, $\epsilon_1 > 0$, and $g_a(t) = (1+\delta_t)^2\log(t)$, $\mathrm{KL}_{\inf}$-UCB2 satisfies*

$$\limsup_{T\to\infty} \frac{\mathbb{E}(N_a(T))}{\log(T)} \leq \frac{1+\tilde{\eta}}{\mathrm{KL}^{\epsilon_1}_{\inf}(\mu_a, m(\mu_1))}.$$

In Lemma 10, Appendix A, we show that $\mathrm{KL}^{\mathcal{L}_B}_{\inf}$ is a continuous function of the moment bound B. Hence, by choosing ϵ_1 close to 0, $\mathrm{KL}_{\inf}$-UCB2's regret gets arbitrarily close to the lower bound, as $T \to \infty$ (modulo $(1+\tilde{\eta})$ factor).

3.2. Conjecture and open problem of Cappé et al. (2013b):

The analysis of the proposed algorithm can be specialized to bound the regret of the Empirical KL-UCB algorithm of Cappé et al. (2013b) for arm-distributions with bounded support. In this

setting, $\mathcal{L} = \mathcal{P}([0,1])$, the collection of all probability measures supported on $[0,1]$. For $\mu_a \in \mathcal{L}$, for each arm $a \in [K]$, and $x \in [0,1]$, upon setting

$$\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, x) \;=\; \inf \;\; \mathrm{KL}(\mu_a, \kappa) \quad \text{s.t.} \quad \kappa \in \mathcal{P}([0,1]), \;\; m(\kappa) \geq x, \tag{6}$$

in the index of our algorithm, we recover the Empirical KL-UCB algorithm.

Proposition 4 *Let $\mathcal{L} = \mathcal{P}([0,1])$. Empirical KL-UCB, with $g_a(t) := \log t + \log\log t$ bounds the pull counts for suboptimal arms $a > 1$ at $T \geq K$ by*

$$\mathbb{E}\left(N_a(T)\right) \;\leq\; \frac{\log(T)}{\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, m(\mu_1)) - O\left((\log T)^{-\frac{1}{5}} \left(\log\log T\right)^{\frac{1}{5}}\right)} + O\left((\log T)^{\frac{4}{5}} \left(\log\log(T)\right)^{\frac{1}{5}}\right).$$

Honda and Takemura (2010) develop an explicit representation and prove properties for the functional defined in (6) above, which we review in Appendix E. Carefully using these in our analysis, we get the above mentioned bound. We refer the reader to Appendix E for the exact bound and its proof.

3.3. Regret analysis

In this section, we prove Theorem 1. The proof of Theorem 3 is similar, and is given in Appendix F.

Henceforth, we assume that arm 1 is the unique arm with maximum mean. The proof proceeds by analysing the events leading to the selection of a sub-optimal arm $a > 1$ by Algorithm 1. We show that if arm $a > 1$ has been sampled enough, then the probability of it getting selected is extremely small. In particular, this corresponds to showing 2 things: first, the $\mathrm{KL}_{\inf}$-UCB index is a high probability upper bound on the true mean, and second, the probability of it being too large is small.

Notice that for $x \in \Re$, and $b \in [K]$, if $N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), x)$ is at least the threshold, $g_b(n)$, then the index for arm b at time n is smaller than x. Similarly, if $N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), x)$ is less than the threshold, then the index computed by $\mathrm{KL}_{\inf}$-UCB is at least x. Proposition 5 below shows that for all n, our index is an upper bound on the true mean of the arm-distribution, with high probability.

Proposition 5 *For $x \geq 0$, $b \in [K]$,*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), m(\mu_b)) - 1 - 2\log(1 + N_b(n)) \geq x\right) \leq e^{-x}.$$

Using Lagrangian duality, it can be shown that $N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), m(\mu_b))$ equals the maximum over the dual variables of a sum of logarithms of i.i.d. random variables with means at most 1. Hence, for fixed dual variables, the exponential of these sum-of-logarithms, which is a product of non-negative random variables with mean at most 1, is a non-negative super-martingale. We construct a mixture of these super-martingales that dominates the exponential of $N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), m(\mu_b))$ after adjusting by $1 + 2\log(1 + N_b(n))$. We refer the reader to Appendix A for a discussion on the dual formulation, and to Section B.1 in the appendix for details of the proof of Proposition 5.

Proof of Theorem 3 also uses a similar concentration inequality for $N_b(n)\,\mathrm{KL}_{\inf}^{\mathcal{L}_{\tilde{B}}}(\hat{\mu}_b(n), m(\mu_b))$ (see Proposition 20), where recall that $\mathcal{L}_{\tilde{B}}$ is the perturbed class used by $\mathrm{KL}_{\inf}$-UCB2. This perturbation helps in getting rid of the additional cost of $2\log(1 + N_b(n))$ incurred above.

We next show that the $\mathrm{KL}_{\inf}$-UCB index being too large is a rare event. Let $\hat{\mu}_{a,s}$ denote the empirical distribution corresponding to s samples from arm a. Lemma 6 below, will be used to bound the probability that the index for a sub-optimal arm $a > 1$ takes value close to the mean of the best-arm, after sufficient samples have been allocated to it.

Lemma 6 *For $\mu \in \mathcal{L}^K$, $0 < \delta < \min_{a \neq 1} \mathrm{KL}_{\inf}(\mu_a, m(\mu_1))$, and $b \in [K]$, there exists $c_\mu > 0$ such that*

$$\mathbb{P}\left(\mathrm{KL}_{\inf}\left(\hat{\mu}_{b,s}, m(\mu_1)\right) \leq \mathrm{KL}_{\inf}\left(\mu_b, m(\mu_1)\right) - \delta\right) \leq e^{-s\, c_\mu \delta^2}.$$

The proof of Lemma 6 above relies on the dual formulation of $\mathrm{KL}_{\inf}$ and its properties, which we review in Appendix A. See Appendix B.2 for the proof of Lemma 6.

3.3.1. PROOF OF THEOREM 1

We now outline the proof of Theorem 1 for the case when $\tilde{\eta} > 0$. When $\tilde{\eta} = 0$, the proof follows similarly with suitable adjustments. Fix $T \geq K+1$. Let $T_j = K + 1 + \sum_{i=K+1}^{j-1} B_i$ denote the random time marking the beginning of the j^{th} batch, where we recall that B_i is the random variable denoting the number of samples to be allocated in the i^{th} batch. In this section, for simplicity of notation, we denote $m(\mu_1)$ by m. The event that at the beginning of the j^{th} batch, a sub-optimal arm a has the maximum index, i.e., $\left\{A_{T_j} = a\right\}$ for $a \neq 1$, is contained in

$$\left\{U_1(T_j) \leq m \quad \text{and} \quad A_{T_j} = a\right\} \bigcup \left\{U_a(T_j) > m \quad \text{and} \quad A_{T_j} = a\right\}, \tag{7}$$

where the left event corresponds to the index for arm 1 evaluating smaller than its true mean at time T_j, while the right one corresponds to the index for the sub-optimal arm taking values higher than the mean of the optimal arm.

Let N be the random number of batches allocated by the algorithm till time T. Recall that the initial K batches correspond to each arm being pulled once. Then, $N_a(T)$ equals $1 + \sum_{j=1}^{N} B_j \mathbb{1}\left(A_{T_j} = a\right)$, and

$$\mathbb{E}\left(N_a(T)\right) = \; 1 + \mathbb{E}\left(D_N\right) + \mathbb{E}\left(E_N\right),$$

where, using the division from (7), we define D_N and E_N as follows:

$$D_N := \sum_{j=K+1}^{N} B_j \mathbb{1}\left(U_1(T_j) \leq m,\; A_{T_j} = a\right), \text{ and } E_N := \sum_{j=K+1}^{N} B_j \mathbb{1}\left(U_a(T_j) > m,\; A_{T_j} = a\right).$$

Let us now look at the deviation of arm a, which will contribute to the dominant term in regret.

Controlling the deviations of sub-optimal arm-$\mathbb{E}\left(E_N\right)$: From the definition of the index of the algorithm, for $t \geq K+1$ and $x \in \Re$, the event $\{U_a(t) \geq x\}$ equals $\{N_a(t)\, \mathrm{KL}_{\inf}(\hat{\mu}_a(t), x) \leq g_a(t)\}$. Fix $\delta > 0$ satisfying $\min_{a>1} \mathrm{KL}_{\inf}\left(\mu_a, m\right) \geq \delta$. Clearly, $\mathbb{1}\left(U_a(T_j) \geq m,\; A_{T_j} = a\right)$ is dominated by the sum of E_{1j} and E_{2j} defined below:

$$E_{1j} = \mathbb{1}\left(\mathrm{KL}_{\inf}\left(\hat{\mu}_a(T_j), m\right) \leq \frac{g_a(T_j)}{N_a(T_j)},\; \mathrm{KL}_{\inf}\left(\hat{\mu}_a(T_j), m\right) > \mathrm{KL}_{\inf}(\mu_a, m) - \delta,\; A_{T_j} = a\right),$$

$$E_{2j} = \mathbb{1}\left(\mathrm{KL}_{\inf}\left(\hat{\mu}_a(T_j), m\right) \leq \mathrm{KL}_{\inf}(\mu_a, m) - \delta,\; A_{T_j} = a\right).$$

Whence,

$$E_N \leq \sum_{j=1}^{N} B_j E_{1j} + \sum_{j=1}^{N} B_j E_{2j}. \tag{8}$$

We argue that E_{1j}, summed over all the batches till time T, contributes the first order term in the regret. Clearly, it is dominated by $\mathbb{1}\left(N_a(T_j)\left(\mathrm{KL}_{\inf}\left(\mu_a, m\right) - \delta\right) \leq g_a(T_j), A_{T_j} = a\right)$, giving

$$\sum_{j=1}^{N} B_j E_{1j} \leq \sum_{j=1}^{N} B_j \mathbb{1}\left(N_a(T_j) \leq \frac{g_a(T_j)}{\mathrm{KL}_{\inf}\left(\mu_a, m\right) - \delta}, A_{T_j} = a\right).$$

Lemma 14 in Appendix B.3 essentially bounds the r.h.s. above, giving

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

The exact form of the $O(\log\log(T))$ term above is given in the proof of Lemma 14.

Next, for $k \geq 2$, let T_a^k denote the random time of the beginning of the batch when arm a won for the k^{th} time. In particular, arm a has been sampled for $(k-1)$ batches till this time. Thus, T_a^k is at least $K - 1 + 1 + \tilde{\eta} + \cdots + \tilde{\eta}(1+\tilde{\eta})^{k-3} = K - 1 + (1+\tilde{\eta})^{k-2}$. Moreover, let $N_{B,a}$ denote the total number of batches allocated to arm a till time T. The other term in (8) satisfies

$$\mathbb{E}\left(\sum_{j=1}^{N} B_j E_{2j}\right) = \mathbb{E}\left(\sum_{k=2}^{N_{B,a}} B_{T_a^k} \mathbb{1}\left(\mathrm{KL}_{\inf}(\hat{\mu}_a(T_a^k), m) \leq \mathrm{KL}_{\inf}(\mu_a, m) - \delta\right)\right).$$

Clearly, $N_a(T_a^k)$ is deterministic. For $k \geq 2$ and $\tilde{\eta} > 0$, it is at least $(1+\tilde{\eta})^{k-1}$, and at most $((1+\tilde{\eta})^{k-1})\tilde{\eta}^{-1}$. Lemma 6 bounds the expectation of the indicator random variable in the above expression. Lemma 15 in Appendix B.3 shows that the bound in this case is proportional to $(1+\tilde{\eta})/(c_\mu \delta^2)$, where c_μ is the bandit instance-dependent constant from Lemma 6. Thus,

$$\mathbb{E}\left(E_N\right) \leq (1+\tilde{\eta})\left(\frac{\log(T)}{\mathrm{KL}_{\inf}\left(\mu_a, m\right) - \delta} + \frac{o(1)}{c_\mu \delta^2} + O(\log\log(T))\right), \tag{9}$$

where the $O(\log\log T)$ terms in the above expression correspond to those in Lemma 14, and the $o(1)$ term is specified in Lemma 15.

Controlling the downward deviation of the optimal arm-$\mathbb{E}\left(D_N\right)$: This term only contributes a constant to the regret till time T. We refer the reader to Lemma 18 in Appendix B.4 for a proof.

Combining everything, we get

$$\mathbb{E}\left(N_a(T)\right) \leq (1+\tilde{\eta})\left(\frac{\log(T)}{\mathrm{KL}_{\inf}\left(\mu_a, m\right) - \delta} + \frac{o(1)}{c_\mu \delta^2} + O(\log\log(T))\right), \tag{10}$$

where the $O(\log\log T)$ terms in the above expression correspond to those in Lemma 14 plus the constant from Lemma 18, and the $o(1)$ term is specified in Lemma 15. The above bound can be optimized over δ. Setting δ to $(c'_\mu(\mathrm{KL}_{\inf}(\mu_a, m(\mu_1)))^2/\log T)^{1/3}$, where $c'_\mu = 2o(1)/c_\mu$ we get that $\mathbb{E}\left(N_a(T)\right)$ is at most

$$(1+\tilde{\eta})\left(\frac{\log T}{\mathrm{KL}_{\inf}(\mu_a, m)} + \frac{3(\log T)^{2/3}(c'_\mu)^{1/3}}{2(\mathrm{KL}_{\inf}(\mu_a, m))^{4/3}} + O((\log T)^{1/3}) + O(\log\log(T))\right), \tag{11}$$

where the $O(\log\log T)$ terms in the above expression correspond to those in Lemma 14, and $O((\log T)^{1/3})$ corresponds to $((\log T)^{1/3}\delta^2/(\mathrm{KL}_{\inf}(\mu_a, m))^3)\sum_i(\delta/\mathrm{KL}_{\inf}(\mu_a, m))^i$.

3.4. Comparison with Robust-UCB with truncated empirical-mean (Bubeck et al., 2013)

It is well known that the standard empirical mean does not concentrate fast when the underlying distributions are heavy-tailed. Bubeck et al. (2013) review three different estimators that concentrate exponentially fast, and propose a UCB algorithm based on these. In this section, for $0 < \epsilon < 1$, we compare our algorithm with that based on the truncated empirical mean estimator (referred to as Robust-UCB, in this section), at the level of confidence intervals around the true mean. The other two estimators proposed by Bubeck et al. (2013) are under different assumptions on the arm-distributions, and are not directly comparable.

Fix $\delta > 0$. Let $\eta \in \mathcal{L}$, and let $\hat{\eta}_n$ denote the empirical distribution based on n samples from η. Recall from (5) that our index is $U_\eta(n) = \max_{\kappa \in \mathcal{L}} \{\mathbb{E}_\kappa(X) : n\,\mathrm{KL}(\hat{\eta}_n, \kappa) \leq C\}$, where C is an appropriately chosen threshold to ensure that $U_\eta(n)$ is an upper bound on $m(\eta)$ with probability at least $1 - \delta$.

The Donsker-Varadhan variational representation for KL-divergence expresses the KL-divergence between any two probability measures P, Q, defined on a common space Ω, as $\mathrm{KL}(P, Q) = \sup_g \left\{\mathbb{E}_P(g(X)) - \log \mathbb{E}_Q\left(e^{g(X)}\right)\right\}$, where the supremum is taken over all measurable functions $g : \Omega \to \Re$ for which $\mathbb{E}_Q\left(e^{g(X)}\right)$ is well-defined. Using this to bound $\mathrm{KL}(\hat{\eta}_n, \kappa)$ in our index, with $g(X) := -\theta X \mathbb{1}(|X| \leq u_n)$, where $u_n = (Bn/\log(\delta^{-1}))^{1/(1+\epsilon)}$, and $\theta > 0$, we get that our index $U_\eta(n)$ is at most

$$\max\left\{\mathbb{E}_\kappa(X) : \kappa \in \mathcal{L} \text{ and } -\theta \sum_{i=1}^{n} X_i \mathbb{1}(|X| \leq u_n) - n \log \mathbb{E}_\kappa\left(e^{-\theta X \mathbb{1}(|X| \leq u_n)}\right) \leq C\right\}. \quad (12)$$

Since $|X\mathbb{1}(|X| \leq u_n)| \leq u_n$, and $\kappa \in \mathcal{L}$, we have $\mathbb{E}_\kappa\left(X^2 \mathbb{1}(|X| \leq u_n)\right) \leq B u_n^{1-\epsilon}$. Let

$$\hat{m}_{1T} := \frac{1}{n} \sum_i X_i \mathbb{1}(|X_i| \leq u_n).$$

Using the standard analysis of Bernstein's inequality to optimize over θ in (12), and substituting for $u_n = (Bn/\log(\delta^{-1}))^{1/(1+\epsilon)}$, we get the following upper bound on our index:

$$\hat{m}_{1T} + B^{\frac{1}{1+\epsilon}} \left(\frac{\log \delta^{-1}}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \left(1 + \left(e^{\frac{C}{\log \delta^{-1}}} - 1\right) \frac{\log \delta^{-1}}{C}\right). \quad (13)$$

We refer the reader to Appendix C for a proof of the above statement. When $C = \log \delta^{-1}$, the above bound on our index is at most the index of the Robust-UCB algorithm (where we note that (13) has improved constants).

We can compare more precisely by considering the application of these indices within the UCB template. In the Robust-UCB algorithm, δ is set to t^{-2}, while we have $\delta = t^{-1}$. This is due to our making use of anytime concentration in Proposition 5, which essentially avoids one union bound in the analysis. We set $C = \log(t) + 2\log N_a(t) + 2\log\log(t)$ in our algorithm. For sub-optimal arms, since $N_a(t) = O(\log(t))$, $C = \log(t) + d\log\log(t)$, for some constant d. In this case, $C/\log(t) = O(1)$, and the bound in (13) recovers the index of Robust-UCB, which is larger than ours. On the other hand, for the optimal-arm, $C \approx 3\log t$, since the number of pulls of the optimal arm is linear. In this case, the Robust-UCB index may be smaller than ours. But this is harmless, as a larger index for arm 1 only increases its chances of being selected.

Taking a step back, we see that concentration inequalities are typically proved starting from an application of Chernoff (including Bernstein/Hoeffding) to get a high probability bound on the moment-generating function of some variable of interest (which may include clipping, truncating or soft versions thereof). From there the consequences for the mean are worked out. In our approach we start instead from the event that κ is in a KL ball around the empirical distribution $\hat{\eta}$. By means of Donsker-Varadhan, this provides MGF control over *all* such variables simultaneously, at the price of the slightly elevated threshold C on the MGF instead of the special-purpose $\ln \delta^{-1}$. Modulo this difference, we hence find that $\mathrm{KL}_{\inf}$-based indices dominate *all* MGF-based indices simultaneously.

We finally remark that the truncated empirical mean estimator considered above is a minor variation of the one proposed by Bubeck et al. (2013), who truncate each sample X_i at a different truncation level u_i, call this $\hat{m}_{2T}$. However, repeating the analysis of Bubeck et al. (2013, Lemma 1) using $\hat{m}_{1T}$ defined above, we in fact obtain tighter confidence intervals, and hence a tighter regret upper-bound (see Bubeck et al. (2013, Proposition 1)). There also does not seem to be a computational advantage to using $\hat{m}_{2T}$ in the context of UCB, as the threshold u_i used for sample X_i depends on the employed confidence level δ, which in turn depends on the current time t, precluding the option of maintaining $\hat{m}_{2T}$ incrementally. Numerically, we observe that Robust-UCB with these two different estimators suffers similar regret. Later, in Section 4, we numerically establish the superiority of our algorithm compared to Robust-UCB.

3.5. The computational cost of $\mathrm{KL}_{\inf}$-UCB

In this section, we discuss the trade-off between the computational cost and statistical optimality of the proposed algorithm. We observe numerically that the time for computing $\mathrm{KL}_{\inf}\left(\hat{\mu}_a(n), \cdot\right)$, increases with $N_a(n)$. This can be seen from its dual formulation (Lemma 7), where η corresponds to the empirical distribution, which grows by one atom every time arm a is sampled. Hence, computing the index for each arm at time n has a cost that is linear in n. Let this linear cost be $c_1 + c_2 n$, where $c_1 \in \Re$ and $c_2 > 0$ are constants.

If the batch size is a constant, say 1 (i.e., when $\tilde{\eta} = 0$), then at each time step the algorithm evaluates the index for each arm, and plays the one with maximum value of the computed index. The total cost of computation for n trials is given by $\sum_{i=1}^{n}(c_1 + c_2 i)$, which is quadratic in n, the total number of trials. For $\tilde{\eta} > 0$, from Theorem 1, each suboptimal arm has at most $d(1+\tilde{\eta})\log n$ samples at time n, for some constant d, while the optimal arm has close to n samples. The number of batches allocated to each sub-optimal arm a till time n, $N_{B,a}(n)$, satisfies

$$(1+\tilde{\eta})^{N_{B,a}(n)+1} \leq d(1+\tilde{\eta})\log(n) \implies N_{B,a}(n) = O\left(\log\log(n)\right).$$

Similar computation gives that $N_{B,1}(n)$, is $O(\log(n))$.

The computational cost for the batches when the best arm won is at most $(1+\tilde{\eta})^{N_{B,1}(n)-1} + N_{B,1}(n)O(\log(n))$, which is $O(n)$. The first term in the previous expression is the total cost of computing the index of arm 1 over the $N_{B,1}(n)$ many batches in which arm 1 won, while the second term is an upper bound on that for the sub-optimal arms. This cost for the batches when sub-optimal arms win is at most $N_{B,a}(n)KO(n) + O(\log(n))$, where the first term is the computational cost of the index of arm 1, which has $O(n)$ samples, contributing $O(n\log\log(n))$ to the cost. Thus, the total worst-case computational cost of $\mathrm{KL}_{\inf}$-UCB with $\tilde{\eta} > 0$ is at most $O(n\log\log(n))$, where the multiplicative constant is given by $\frac{1}{\log(1+\tilde{\eta})}$.

Computing $\mathrm{KL}_{\inf}$-UCB index: Computing $\mathrm{KL}_{\inf}$-UCB index in the original form would require inverting the $\mathrm{KL}_{\inf}$ functional. Since there is no closed form expression for $\mathrm{KL}_{\inf}$, one approach

for solving this can be by binary search for the second argument, computing $\mathrm{KL}_{\inf}$ at each iteration. However, the representation of the index in (5) gives it as the solution to a single optimization problem. This is discussed in Appendix D.

4. Numerical results

In this section we discuss the numerical studies undertaken to demonstrate the superiority of $\mathrm{KL}_{\inf}$-UCB. We run our algorithm on two different bandit problems. In both these experiments, the algorithm is presented with a two-armed bandit, each arm having a Generalized Pareto (GenPar) distribution. It is a heavy-tailed distribution, which has 3 parameters, μ, σ, ζ, which correspond to location, scale, and shape, respectively. When $\zeta > 0$, it has a density function given by $h(x) = \sigma^{-1}\left(1 + \zeta\sigma^{-1}(x-\mu)\right)^{-1-\zeta^{-1}}$, for $x \geq \mu$.

In the first experiment, we let $\epsilon = 0.7$, B is set to 7, and $\mathcal{L}$ is the collection of all distributions with 1.7^{th}-moment bounded by 7. Arm 1 is GenPar$(-1, 2, 0.2)$, and arm 2 is GenPar$(-1, 1, 0.2)$. Hence, arm 1 is optimal with mean $= 1.5$, and whenever the algorithm chooses arm 2, it suffers a regret of 1.25. Moreover, $\mathrm{KL}_{\inf}(\mu_2, 1.5) \approx 0.1$, which is evaluated with respect to the class $\mathcal{L}$.

We compare the performance of $\mathrm{KL}_{\inf}$-UCB (with $\tilde{\eta}$ set to 0.1) to the asymptotic lower bound, both with the theoretical threshold of Theorem 1 and with an agressive, practically-tuned threshold of $\log t$, which is even smaller than that for $\mathrm{KL}_{\inf}$-UCB2 (Theorem 3). In Figure 1, we plot the regret incurred by both of these, the asymptotic lower bound, and 1.1 times the lower bound, which is the asymptotic upper-bound for $\mathrm{KL}_{\inf}$-UCB with $\tilde{\eta} = 0.1$ (Corollary 2).

As can be seen from Figure 1, the algorithm with the aggressive threshold performs significantly better. It, in fact, suffers regret lower than the lower bound, even for large horizons. This is not surprising, and highlights the asymptotic nature of the lower bound. In $\mathrm{KL}_{\inf}$-UCB on the other hand, the $\log\log t$ term contributes significantly to the threshold over the horizon considered, leading to higher regret. Henceforth, we only consider $\mathrm{KL}_{\inf}$-UCB with threshold set to $\log(t)$.

We also compare the performance of the agressive algorithm with that of Robust-UCB with the truncation based estimator, proposed by Bubeck et al. (2013). Figure 2 plots the ratio of regret suffered by Robust-UCB and our algorithm, on two different bandit instances. The "Easy problem" in the figure corresponds to the set-up of experiment 1 described above.

In the second experiment, we consider a slightly more difficult-to-learn setting, where the tails of the arm-distribution are heavier than in the first experiment. We see that both the algorithms suffer more regret on this problem compared to the previous bandit instance. However, the performance of Robust-UCB degrades much more. For this experiment, we set ϵ to 0.1, and let $B = 13$, whence, $\mathcal{L}$ is collection of all distributions with 1.1^{th}-moment bounded by 13. The arm-distributions are set to GenPar$(2.17, 3.7, 0.5)$ for arm 1, and GenPar$(-1, 2, 0.71)$ for arm 2. In this setting, the optimal arm has mean 9.57, and when the sub-optimal arm is pulled, the algorithm suffers 3.674 units of regret.

Figure 2 shows that even with our batching, we significantly out-perform the Robust-UCB algorithm. It suffers retret that is 40 times that of $\mathrm{KL}_{\inf}$-UCB on the easy problem (setting of experiment 1), and the regret ratio is 19 the difficult setting of experiment 2.

5. Conclusion

We consider minimising regret for heavy-tailed bandits. Our approach follows the UCB template. But instead of constructing upper confidence intervals, we "let the lower bound speak" and end up

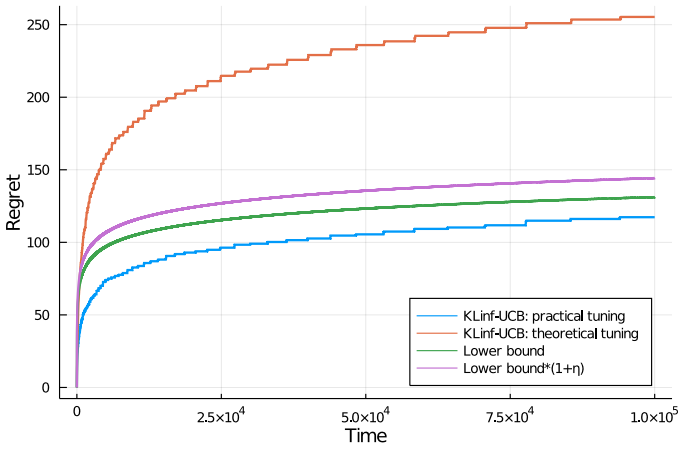

Figure 1: Comparison of regret of our algorithm with different thresholds, with batch-multiplicative factor set to $\tilde{\eta} = 0.1$ in each, the asymptotic lower bound of Burnetas and Katehakis (1996) and the asymptotic upper bound in Theorem 1. Plots are averaged over 100 independent experiments. The batches are visible in the staircase pattern: long horizontal steps correspond to batches in which the optimal arm is chosen, and each vertical step represents a short batch where a suboptimal arm is picked.

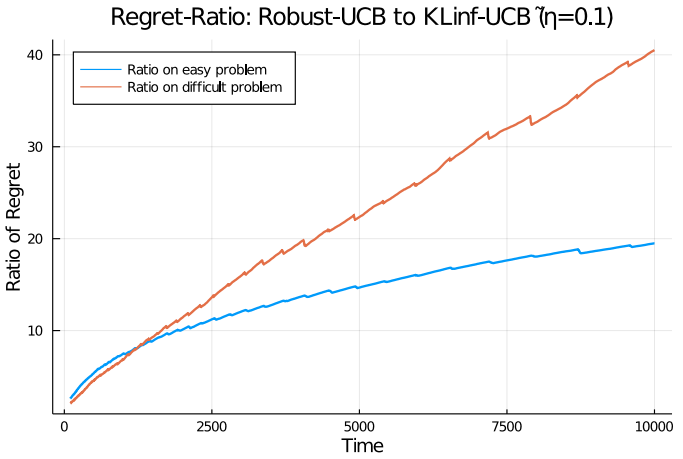

Figure 2: Ratio of regret of Robust-UCB with truncated empirical mean and $\mathrm{KL}_{\mathrm{inf}}$-UCB with batch-multiplicative factor set to $\tilde{\eta} = 0.1$ and an aggressive threshold of $\log t$, for simple and difficult bandit instances. This figure demonstrates that the performance of Robust-UCB degrades much more than that of our algorithm on a difficult bandit instance.

with $\mathrm{KL}_{\inf}$-based indices. We exploit the dual characterization of $\mathrm{KL}_{\inf}$ and the associated index, which guides both our index implementation and its statistical concentration analysis. We show that with these ingredients it is possible to match the instance-optimal regret lower bounds. Using a batched sampling scheme, we achieve total run time almost linear in the number of samples, at the cost of a small constant factor in the regret. We prove that $\mathrm{KL}_{\inf}$ indices dominate the UCB index based on the standard mean estimators for heavy-tailed distributions. We empirically validate that this translates into much improved regret on synthetic problems. Here are a few remarks to conclude the paper

- One may generalise the moment constraint $\mathbb{E}\left(|X|^{1+\epsilon}\right)$ by requiring instead that $\mathbb{E}\left(f(X)\right) \leq B$ for some convex (and super-linear) function f. To make this work, one would have to prove compactness of the corresponding region for $\boldsymbol{\lambda}$ in the dual formulation, so as to make the uniform-prior based regret analysis work. Or one would have to prove continuity of $\mathrm{KL}_{\inf}$ in its second argument and the parameters of the class to employ the perturbation-based approach from Appendix F.

- We see in experiments that the performance is sensitive to the choice of threshold, and that our theoretically motivated thresholds are (currently) conservative. Our thresholds come from mixture martingales with universal coding/regret guarantees. Perhaps a redundancy-based analysis would be better suited for proving tighter concentration inequalities, and could reduce the threshold from $\log n$ to $\log \log n$.

- We considered the class of distributions with bounded *uncentered* $(1+\epsilon)^{\text{th}}$ moment. A natural problem would be to extend the approach to the centred analogue. Our approach relies on the dual formulation of $\mathrm{KL}_{\inf}$. However, bounded centered-moment is no longer a convex constraint in the distribution, rendering the class $\mathcal{L}$, and hence the corresponding $\mathrm{KL}_{\inf}$ optimization problem non-convex. Handling this non-convexity would require development of more nuanced techniques.

- Suppose that the samples from an arm are no-longer i.i.d., but satisfy only the milder condition that their conditional mean is fixed, and the conditional $(1+\epsilon)^{th}$-moment is bounded by B (we may think of this as an imprecise probability analogue of the i.i.d. problem). As discussed in Remark 12 in Section B.1, our martingale-based proof for concentration of $\mathrm{KL}_{\inf}$ (Proposition 5) is valid even with this relaxation, giving a high-probability confidence interval for the fixed mean. However, it is not clear what a regret-minimization algorithm, or for that matter an expected regret lower bound would look like in this setting.

Acknowledgments

We acknowledge the support of the Department of Atomic Energy, Government of India, to TIFR under project no. RTI4001.

References

Rajeev Agrawal. Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shubhada Agrawal, Sandeep Juneja, and Peter Glynn. Optimal δ-correct best-arm selection for heavy-tailed distributions. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 61–110. PMLR, 08 Feb–11 Feb 2020a.

Shubhada Agrawal, Wouter M. Koolen, and Sandeep Juneja. Optimal best-arm identification methods for tail-risk measures. *arXiv preprint arXiv:2008.07606*, 2020b.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, May 2002.

C. Berge. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity*. Dover books on mathematics. Dover Publications, 1997. ISBN 9780486696539.

P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley, 2013.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tails. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

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

O Cappé, A Garivier, OA Maillard, R Munos, and G Stoltz. Kullback–leibler upper confidence bounds/supplemental article. *Ann. Stat*, 41(3):1516–1541, 2013a.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013b.

Wesley Cowan and Michael N Katehakis. An asymptotically optimal policy for uniform bandits of unknown support. *arXiv preprint arXiv:1505.01918*, 2015.

Wesley Cowan, Junya Honda, and Michael N. Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

Amir Dembo and Ofer Zeitouni. Large deviations techniques and applications. corrected reprint of the second (1998) edition. stochastic modelling and applied probability, 38, 2010.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *In Proceedings of the Twenty-third Conference on Learning Theory (COLT 2010*, pages 67–79. Omnipress, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *The Journal of Machine Learning Research*, 16(1):3721–3756, 2015.

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

Prashanth L.A., Krishna Jagannathan, and Ravi Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5577–5586. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/l-a-20a.html.

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

Tor Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. In *Advances in Neural Information Processing Systems*, pages 1584–1593, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514, 2011.

João Nicolau and Paulo MM Rodrigues. A new regression-based tail index estimator. *Review of Economics and Statistics*, 101(4):667–680, 2019.

E. Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58 (5):527–535, 09 1952.

Rangarajan K. Sundaram. *A First Course in Optimization Theory*. Cambridge University Press, 1996.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Sattar Vakili, Keqin Liu, and Qing Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5): 759–767, 2013.

Appendix A. $\mathrm{KL}_{\inf}$ - dual formulation and properties

In this appendix, we review some background and results that will be useful in the analysis of the proposed algorithms.

Lévy metric and weak convergence: The Lévy metric, $d_L(\kappa, \eta)$, between probability distributions (see Dembo and Zeitouni (2010, Appendix D)) κ and η on $\Re$, is given by

$$d_L(\eta, \kappa) = \inf \{\zeta > 0 : F_\kappa(x-\zeta) - \zeta \leq F_\eta(x) \leq F_\kappa(x+\zeta) + \zeta, \ \forall x \in \Re\},$$

where for $\psi \in \mathcal{P}(\Re)$, F_ψ denotes the CDF function for ψ, i.e., for $x \in \Re$, $F_\psi(x) := \psi(-\infty, x]$. Furthermore, weak convergence of sequences of probability measures is equivalent to their convergence in the Lévy metric (see Billingsley (2013, Theorem 6.8), and Dembo and Zeitouni (2010, Theorem D.8)).

Next, let us look at the functional that appears in the lower bound on expected regret. For $B > 0$ and $\epsilon > 0$, recall that $\mathcal{L}_B = \left\{\kappa \in \mathcal{P}(\Re) : \mathbb{E}_\kappa\left(|X|^{1+\epsilon}\right) \leq B\right\}$. Let $x \in \Re$ be such that $|x|^{1+\epsilon} < B$, and for $\eta \in \mathcal{P}(\Re)$, recall that $m(\eta) := \mathbb{E}_\eta(X)$, and

$$\mathrm{KL}_{\inf}^{\mathcal{L}_B}(\eta, x) = \inf \{\mathrm{KL}(\eta, \kappa) : \kappa \in \mathcal{L}_B, \ m(\kappa) \geq x\}.$$

In the rest of the appendix, we denote $\mathrm{KL}_{\inf}^{\mathcal{L}_B}(\eta, x)$ by $\mathrm{KL}_{\inf}(\eta, x)$, and $\mathcal{L}_B$ by $\mathcal{L}$. Furthermore, let $\mathrm{Supp}(\eta)$ denote the support of measure η. We first review an alternative representation and properties of the function $\mathrm{KL}_{\inf}$, which will be useful in the analysis of our algorithm. The following Lemma is due to Agrawal et al. (2020a, Theorem 12), which we state for completeness.

Lemma 7 (Dual formulation of $\mathrm{KL}_{\inf}$) *For $\eta \in \mathcal{P}(\Re)$ and x such that $|x|^{1+\epsilon} < B$,*

$$\mathrm{KL}_{\inf}(\eta, x) = \max_{(\lambda_1, \lambda_2) \in \mathcal{R}(x, B)} \mathbb{E}_\eta\left(\log\left(1 - (X-x)\lambda_1 - (B - |X|^{1+\epsilon})\lambda_2\right)\right), \tag{14}$$

where

$$\mathcal{R}(x, B) = \left\{(\lambda_1 \geq 0, \lambda_2 \geq 0) : \frac{\epsilon \lambda_1^{1+\frac{1}{\epsilon}}}{\lambda_2^{\frac{1}{\epsilon}}(1+\epsilon)^{1+\frac{1}{\epsilon}}} + B\lambda_2 - x\lambda_1 - 1 \leq 0\right\}.$$

There is a unique $(\lambda_1^*, \lambda_2^*)$ that achieves the maximum above. Furthermore, any primal variable that achieves the infimum in (14), satisfies

$$\frac{d\kappa^*}{d\eta}(y) = \left(1 - (y-x)\lambda_1^* - (B - |y|^{1+\epsilon})\lambda_2^*\right)^{-1}, \quad \text{for } y \in \mathrm{Supp}(\eta).$$

$|\mathrm{Supp}(\kappa^*) \setminus \mathrm{Supp}(\eta)|$ is at most 1. Furthermore,

$$1 - (y^* - x)\lambda_1^* - (B - |y^*|^{1+\epsilon})\lambda_2 = 0 \quad \text{for } y^* \in \{\mathrm{Supp}(\kappa^*) \setminus \mathrm{Supp}(\eta)\}.$$

Lemma 8 *The region $\mathcal{R}(x, B)$ defined in Lemma 7 is compact whenever $|x|^{1+\epsilon} < B$. It satisfies*

$$0 \leq \lambda_1 \leq \left(B^{\frac{1}{1+\epsilon}} - x\right)^{-1}, \quad \text{and} \quad 0 \leq \lambda_2 \leq \left(B - |x|^{1+\epsilon}\right)^{-1}.$$

Proof Let $h(\lambda_1, \lambda_2) = \epsilon\lambda_1^{1+\frac{1}{\epsilon}} + B\lambda_2^{1+\frac{1}{\epsilon}}(1+\epsilon)^{1+\frac{1}{\epsilon}} - (x\lambda_1 + 1)\lambda_2^{\frac{1}{\epsilon}}(1+\epsilon)^{1+\frac{1}{\epsilon}}$. Then the constraint in $\mathcal{R}(x, B)$ is $h(\lambda_1, \lambda_2) \leq 0$. Clearly, h is convex in λ_1 and λ_2. Thus, the given constraint implies $\min_{\lambda_1 \geq 0} h(\lambda_1, \lambda_2) \leq 0$. This gives the bound on λ_2. Similarly, $\min_{\lambda_2 \geq 0} h(\lambda_1, \lambda_2) \leq 0$ gives the bound on λ_1. ■

For a fixed x, $\mathrm{KL}_{\inf}(\cdot, x)$ is a function from $\mathcal{P}(\Re)$ to $\Re^+$. Recall that we endow the space $\mathcal{P}(\Re)$ with the the topology of weak convergence, or equivalently with the Lévy metric (see, e.g., Dembo and Zeitouni (2010, Appendix D) for definitions of the two topologies and their equivalence).

Lemma 9 (Properties of $\mathcal{L}_B$ and $\mathcal{R}(x, B)$) *$\mathcal{L}$ is a a uniformly integrable collection, and a compact set of probability measures in the topology of weak convergence. Furthermore,*

1. *for $B > 0$, $\mathcal{R}(x, B)$ is an upper-hemicontinuous function of x on $(-B^{\frac{1}{1+\epsilon}}, B^{\frac{1}{1+\epsilon}})$,*
2. *for $x \in \Re$, $\mathcal{R}(x, B)$ is an upper-hemicontinuous function of B on $(|x|^{1+\epsilon}, \infty)$.*

Proof Uniform integrability and compactness of $\mathcal{L}$ follow from (Agrawal et al., 2020b, Lemma 3.2).

Proof of 1: To see upper-hemicontinuity of $\mathcal{R}(x, B)$ for a fixed B, consider a sequence $x_n \to x$. Let $\eta_n \in \mathcal{R}(x_n, B)$ be a sequence of measures in $\mathcal{L}$, which is a tight collection of probability measures in the weak topology. Then, there is a subsequence η_{p_i} which converges weakly to η_p (see Billingsley (2013)). From (Sundaram, 1996, Proposition 9.8), it is sufficient to show that η_p belongs to $\mathcal{R}(x, B)$. Clearly, η_p belongs to class $\mathcal{L}$ since it is a closed set, and η_{p_i} belong to $\mathcal{L}_B$. Furthermore, since $\mathcal{L}$ is a uniformly integrable collection, $\eta_n \overset{D}{\Rightarrow} \eta$ implies that $\mathbb{E}_{\eta_{p_i}}(X) \to \mathbb{E}_{\eta_p}(X)$, and hence, $\mathbb{E}_{\eta_p}(X) \geq x$.

Proof of 2: To see upper-hemicontinuity of $\mathcal{R}(x, B)$ for a fixed x, consider a sequence $B_n \to B$, and let η_n be a sequence in $\mathcal{R}(x, B_n)$. For any fixed $\delta > 0$, there exists n_0 such that for all $n \geq n_0$, $B_n \leq B + \delta$, and the sequence $\{\eta_n\}$, for all $n \geq n_0$, satisfies $\mathbb{E}_{\eta_n}\left(|X|^{1+\epsilon}\right) \leq B + \delta$. Let $\mathcal{L}_{B+\delta}$ denote the collection of all probability measures with $(1+\epsilon)^{th}$-moment bounded by $B + \delta$. As above, since $\mathcal{L}_{B+\delta}$ is a closed, tight, and uniformly integrable collection of probability measures, arguments as in the previous paragraph show that $\mathbb{E}_{\eta_{p_i}}(X) \to \mathbb{E}_{\eta_p}(X) \geq x$, and η_p belongs to $\mathcal{L}_{B+\delta}$. Since δ was arbitrary, η_p belongs to $\mathcal{L}$, and hence to $\mathcal{R}(x, B)$. ■

Lemma 10 (Properties of $\mathrm{KL}_{\inf}$) *For a fixed $\eta \in \mathcal{P}(\Re)$ and*

1. *for a fixed $B > 0$, $\mathrm{KL}_{\inf}(\eta, x)$ is a continuous function of x on $\left(-B^{\frac{1}{1+\epsilon}}, B^{\frac{1}{1+\epsilon}}\right)$.*
2. *for a fixed $x \in \Re$, $\mathrm{KL}_{\inf}(\eta, x)$ is a continuous function of B on $\left(|x|^{1+\epsilon}, \infty\right)$.*

Proof To prove the continuity in x and B, we prove lower- and upper-semicontinuity separately. To prove 1, we first argue that for a fixed B, $\mathrm{KL}_{\inf}(\eta, x)$ is a convex function of x, and hence an upper-semicontinuous function.

Next, for $y \in \Re$, we define $f(y) := |y|^{1+\epsilon}$, and for $y \in [0, B]$, $f^{-1}(y) := y^{\frac{1}{1+\epsilon}}$. To see the convexity, consider $x_1, x_2 \in \left(-f^{-1}(B), f^{-1}(B)\right)$. Let

$$\mathcal{R}(x, B) := \{\gamma \in \mathcal{P}(\Re) : \mathbb{E}_\gamma(X) \geq x,\ \mathbb{E}_\gamma(f(X)) \leq B\}.$$

Clearly, the set $\mathcal{R}(x, B)$ is convex and non-empty for the choices of x and B under consideration. Let $\kappa_1, \kappa_2 \in \mathcal{R}(x, B)$ be such that $\mathrm{KL}_{\inf}(\eta, x_1) = \mathrm{KL}(\eta, \kappa_1)$ and $\mathrm{KL}_{\inf}(\eta, x_2) = \mathrm{KL}(\eta, \kappa_2)$. Existence of κ_1 and κ_2 is guaranteed by compactness of $\mathcal{R}(x, B)$ and lower-semicontinuity of KL. Furthermore, for $\lambda \in (0, 1)$,

$$R_{12}(\lambda) := \{\lambda \mathcal{R}(x_1, B) + (1-\lambda)\mathcal{R}(x_2, B)\} \subset \mathcal{R}(\lambda x_1 + (1-\lambda)x_2, B).$$

Consider the following inequalities:

$$\mathrm{KL}_{\inf}(\eta, \lambda x_1 + (1-\lambda)x_2) \ \leq \ \inf_{\kappa \in R_{12}(\lambda)} \mathrm{KL}(\eta, \kappa) \ \leq \ \mathrm{KL}(\eta, \lambda\kappa_1 + (1-\lambda)\kappa_2).$$

Using joint convexity of KL, the above can further be bounded from above by $\lambda\,\mathrm{KL}(\eta, \kappa_1) + (1-\lambda)\,\mathrm{KL}(\eta, \kappa_2)$, which equals $\lambda\,\mathrm{KL}_{\inf}(\eta, x_1) + (1-\lambda)\,\mathrm{KL}_{\inf}(\eta, x_2)$ by choice of κ_1 and κ_2, giving

$$\mathrm{KL}_{\inf}(\eta, \lambda x_1 + (1-\lambda)x_2) \leq \lambda\,\mathrm{KL}_{\inf}(\eta, x_1) + (1-\lambda)\,\mathrm{KL}_{\inf}(\eta, x_2).$$

Convexity in B also follows similarly, proving the upper-semicontinuity of $\mathrm{KL}_{\inf}$ in x and in B, under the given conditions.

For $\eta \in \mathcal{P}(\Re)$, since $\mathrm{KL}(\eta, \cdot)$ is a lower-semicontinuous function in the topology of weak convergence (see Posner (1975)), and the region of optimization, $\mathcal{R}(x, B)$, is a non-empty, compact, upper-hemicontinuous correspondence of x and B under the respective given conditions (Lemma 9), the optimal value, $\mathrm{KL}_{\inf}(\eta, x)$ is lower-semicontinuous in x for a fixed B, and lower-semicontinuous in B for a fixed x (see (Berge, 1997, Theorem 2, page 116)). ■

Appendix B. Results related to regret guarantees

In this appendix, we state and prove the results that assist us in the proof of Theorem 1.

B.1. Towards proving Proposition 5

In this section, we prove the anytime concentration inequality in Proposition 5. The proof involves constructing mixtures of super-martingales using the dual formulation for $\mathrm{KL}_{\inf}$, and may also be of independent interest. Lemma 11 below is borrowed from (Agrawal et al., 2020b, Lemma E.1), and is stated here for completeness.

Lemma 11 *Let $\Lambda \subseteq \mathbb{R}^d$ be a compact and convex subset and let q be the uniform distribution on Λ. Let $g_i : \Lambda \to \mathbb{R}$ be any series of exp-concave functions. Then*

$$\max_{\boldsymbol{\lambda} \in \Lambda} \sum_{i=1}^{T} g_i(\boldsymbol{\lambda}) \ \leq \ \log \mathbb{E}_{\boldsymbol{\lambda} \sim q}\left(e^{\sum_{i=1}^{T} g_t(\boldsymbol{\lambda})}\right) + d\log(T+1) + 1.$$

B.1.1. PROOF OF PROPOSITION 5

We first note that if $\mu_b = \delta_{B^{1/(1+\epsilon)}}$, which is the only distribution in $\mathcal{L}_B$ with the maximum possible mean in this class, i.e., $B^{1/(1+\epsilon)}$, then the statement is vacuously true, since $\hat{\mu}_b(n) = \mu_b$. Whence $\mathrm{KL}_{\inf}(\hat{\mu}_b, m(\mu_b)) = 0$.

We now consider $\mu_b \neq \delta_{B^{1/(1+\epsilon)}}$. For $\epsilon > 0$ and $B > 0$, let $f(y) := |y|^{1+\epsilon}$, and for $c \in [0, B]$ define $f^{-1}(c) = c^{1/(1+\epsilon)}$. From the dual formulation in Lemma 7,

$$N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), m(\mu_b)) = \max_{\boldsymbol{\lambda} \in \mathcal{R}(m(\mu_b), B)} \sum_{i=1}^{N_b(n)} \log\left(1 - \lambda_1(X_i - m(\mu_b)) - \lambda_2(B - f(X_i))\right),$$

where $\mathcal{R}(m(\mu_b), B) \subset \Re^2$. Let q be uniform distribution on $\mathcal{R}(m(\mu_b), B)$, which is defined since the region is a compact set (see Lemma 8). Define

$$U_b(n) = \mathbb{E}_{\boldsymbol{\lambda} \sim q}\left(\prod_{i=1}^{N_b(n)} \left(1 - \lambda_1(X_i - m(\mu_b)) - \lambda_2(B - f(X_i))\right) \Big| X_1, \ldots, X_{N_b(n)}\right) \text{ for } n \geq 1,$$

where $\left\{X_1, \ldots, X_{N_b(n)}\right\}$ are $N_b(n)$ samples generated from arm b in time n. Setting $d = 2$, $g_i(\boldsymbol{\lambda}) = \log(1 - \lambda_1(X_i - m(\mu_b)) - \lambda_2(B - f(X_i)))$, in Lemma 11, on each sample path, we have

$$N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), m(\mu_b)) \leq \log U_b(n) + 2\log(1 + N_b(n)) + 1. \tag{15}$$

Since $\mu_b \in \mathcal{L}$, $U_b(n)$ is a non-negative super-martingale with mean at most 1. Using this in (15), together with Ville's inequality, we get

$$\mathbb{P}\left(\exists n \in \mathbb{N} :\ N_b(n)\,\mathrm{KL}_{\inf}(\hat{\mu}_b(n), m(\mu_b)) - 2\log(1 + N_b(n)) - 1 \geq x\right) \leq e^{-x},$$

proving the desired inequality. □

Remark 12 The concentration inequality in Proposition 5 shows that $U_b(n)$ (defined in (5)), with an appropriate choice of threshold, $g_b(n)$, is a high probability upper confidence interval for the true mean for arm b at time n. Taking a step back, let $X_1, \ldots, X_s$ denote s i.i.d. samples from an underlying distribution, $\eta \in \mathcal{L}$, and let the empirical distribution corresponding to these s samples be denoted by $\hat{\eta}_s$. Then, for $\delta > 0$, Proposition 5 shows that

$$U_\eta(s) := \max\ \left\{x :\ s\,\mathrm{KL}_{\inf}(\hat{\eta}_s, x) \leq \log \delta^{-1} + 2\log(s+1) + 1\right\} \tag{16}$$

is an upper bound on true mean of η, $m(\eta)$, with probability at least $1 - \delta$.

However, observe that our martingale-based proof does not rely on the samples being i.i.d. Suppose the samples, $X_1, \ldots, X_s$ only satisfy the following: for $p \in [-B^{\frac{1}{1+\epsilon}}, B^{\frac{1}{1+\epsilon}}]$,

$$\mathbb{E}\left(X_i \middle| X_1, \ldots, X_{i-1}\right) = p, \quad \text{and} \quad \mathbb{E}\left(|X_i|^{1+\epsilon} \middle| X_1, \ldots, X_{i-1}\right) \leq B.$$

Even under this mild condition, (16) gives an upper confidence interval for p with probability at least $1 - \delta$, showing that our $\mathrm{KL}_{\inf}$-based approach is robust to some of the underlying assumptions, and may be used to develop robust statistical learning procedures.

B.2. Proof of Lemma 6

Let us first consider $\mu_b = \delta_{-B^{1/(1+\epsilon)}}$. In this case, the statement is vacuously true since $\hat{\mu}_{b,s} = \mu_b$.

Now, let $\mu_b \neq \delta_{-B^{1/(1+\epsilon)}}$. For $\epsilon > 0$ and $B > 0$ let $f(y) = |y|^{1+\epsilon}$. For $c \in [0, B]$ define $f^{-1}(c) = c^{1/(1+\epsilon)}$. Using the dual formulation for $\mathrm{KL}_{\inf}(\hat{\mu}_{b,s}, m(\mu_1))$ from Lemma 7, the required probability equals

$$\mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\mathcal{R}(m(\mu_1),B)} \frac{1}{s}\sum_{i=1}^{s}\log\left(1-\left(X_i-m(\mu_1)\right)\lambda_1-\left(B-f\left(X_i\right)\right)\lambda_2\right) \leq \mathrm{KL}_{\inf}\left(\mu_b, m(\mu_1)\right)-\delta\right),$$

which is bounded from above by

$$\mathbb{P}\left(\frac{1}{s}\sum_{i=1}^{s}\log\left(1-\left(X_i-m(\mu_1)\right)\lambda_1^*-\left(B-f\left(X_i\right)\right)\lambda_2^*\right) \leq \mathrm{KL}_{\inf}\left(\mu_b, m(\mu_1)\right)-\delta\right),$$

where

$$(\lambda_1^*, \lambda_2^*) \in \underset{\boldsymbol{\lambda}\in\mathcal{R}(m(\mu_1),B)}{\operatorname{argmax}} \mathbb{E}_{\mu_b}\left(\log\left(1-\left(X-m(\mu_1)\right)\lambda_1-\left(B-f\left(X\right)\right)\lambda_2\right)\right).$$

For $\theta \geq 0$, the required probability can be bounded by

$$\mathbb{P}\left(-\theta\sum_{i=1}^{s}\log\left(1-\left(X_i-m(\mu_1)\right)\lambda_1^*-\left(B-f\left(X_i\right)\right)\lambda_2^*\right) \geq -s\theta\left(\mathrm{KL}_{\inf}\left(\mu_b, m(\mu_1)\right)-\delta\right)\right),$$

which can further be bounded by

$$\mathbb{E}_{\mu_b}\left(\exp\left\{-\theta\sum_{i=1}^{s}\log\left(1-\left(X_i-m(\mu_1)\right)\lambda_1^*-\left(B-f\left(X_i\right)\right)\lambda_2^*\right)\right\}\right)e^{s\theta(\mathrm{KL}_{\inf}(\mu_b,m(\mu_1))-\delta)}.$$

Let $Y_i := \log\left(1-\left(X_i-m(\mu_1)\right)\lambda_1^*-\left(B-f\left(X_i\right)\right)\lambda_2^*\right)$. Since $X_i \sim \mu_b$ are i.i.d., Y_i are i.i.d. as well. Furthermore, let Y be independent and identically distributed as Y_i, and for $\gamma \leq 1$, let $\Lambda_Y := \log\mathbb{E}\left(\exp\left\{\gamma Y\right\}\right)$. Observe that $\mathbb{E}_{\mu_b}\left(Y\right) = \mathrm{KL}_{\inf}\left(\mu_b, m(\mu_1)\right)$. Then the above expression equals

$$\exp\left\{s\left(\theta\left(\mathrm{KL}_{\inf}\left(\mu_b, m(\mu_1)\right)-\delta\right)+\Lambda_Y(-\theta)\right)\right\}.$$

Since the previous bound is true for all values of $\theta \geq 0$, in particular, we have that the required probability is bounded by

$$\exp\left\{-s\sup_{\theta\leq 0}\left\{\theta\left(\mathrm{KL}_{\inf}\left(\mu_b, m(\mu_1)\right)-\delta\right)-\Lambda_Y(\theta)\right\}\right\}.$$

Obsesrve that $\Lambda_Y(0) = 0$ and $\Lambda_Y(-1) \leq 0$. To see the latter,

$$\Lambda_Y(-1) = \log\mathbb{E}_{\mu_b}\left(\frac{1}{1-\left(X-m(\mu_1)\right)\lambda_1^*-\left(B-f\left(X\right)\right)\lambda_2^*}\right) = \log\kappa_b(\mathrm{Supp}(\mu_b)) \leq 0,$$

where κ_b is the optimal primal variable for the $\mathrm{KL}_{\inf}(\mu_b, m(\mu_1))$ optimization problem, and satisfies $\kappa_b(\mathrm{Supp}(\mu_b)) \leq 1$ (see Lemma 7). Furthermore, $\Lambda_Y(\gamma)$ is a convex function of γ. Whence, the supremum in

$$\sup_{\theta \leq 0} \{\theta(\mathrm{KL}_{\inf}(\mu_b, m(\mu_1)) - \delta) - \Lambda_Y(\theta)\}$$

is attained at some $\theta^* < 0$, and the optimal value equals $I_Y(\mathrm{KL}_{\inf}(\mu_b, m(\mu_1) - \delta))$, where

$$I_Y(\mathrm{KL}_{\inf}(\mu_b, m(\mu_1)) - \delta) := \sup_{\theta} \{\theta(\mathrm{KL}_{\inf}(\mu_b, m(\mu_1)) - \delta) - \Lambda_Y(\theta)\}$$

is the large deviations rate function for the random variable Y.

It is easy to check that Y satisfies Condition 1 below. Lemma 13 then shows that there exists $\delta_0 > 0$ and a constant c_b, such that for all $\delta < \delta_0$, $I_Y(\mathrm{KL}_{\inf}(\mu_b, m(\mu_1)) - \delta) \geq c_b\delta^2$. Since $\delta \in [0, \min_{a \neq 1} \mathrm{KL}_{\inf}(\mu_a, m(\mu_1))]$, by convexity of I_Y, there exists a constant c'_b (possibly smaller than c_b), such that $I_Y(\mathrm{KL}_{\inf}(\mu_b, m(\mu_1)) - \delta) \geq c'_b\delta^2$. Then, $c_\mu := \min_{a \neq 1} c'_a$. Clearly, $I_Y(\mathrm{KL}_{\inf}(\mu_b, m(\mu_1)) - \delta) \geq c_\mu\delta^2$, giving the desired bound. □

Condition 1: Given a random variable X with $\Lambda_X(\theta) < \infty$ for $\theta \in [-1, 1]$. Let $I_X(x) := \sup_\theta \{\theta x - \Lambda_X(\theta)\}$ denote the associated large deviation rate function. Let $m(X) > 0$, i.e., $\Lambda'_X(0) > 0$, and hence $I(m(X)) = 0$.

Lemma 13 *For X satisfying Condition 1 above, $\exists \delta_0 > 0$ such that $\forall \delta \in [0, \delta_0]$, there exists a constant $c > 0$ such that $I(m(X) - \delta) \geq c\delta^2$.*

Proof Observe that if $\Lambda''_X(0) = 0$, then X is degenerate, giving $I(x) = \infty$ for $x \neq m(X)$. Consider a non-degenerate X, meaning $\Lambda''_X(0) > 0$. Also, consider the θ_y that satisfies $\Lambda'_X(\theta_y) = y$. Here, $\theta_{m(X)} = 0$ and $\theta_y < 0$ for $y < m(X)$. From Condition 1, $\Lambda'_X(\theta)$ is continuously differentiable for $\theta \in (-1, 1)$. By Implicit Function Theorem, there exists a neighbourhood $(m(X) - \tilde{\delta}_0, m(X))$ such that for $y \in (m(X) - \tilde{\delta}_0, m(X))$, θ_y exists and is continuously differentiable.

Next, recall that $\theta_y < 0$ for $y < m(X)$. Choose $\tilde{\delta}_0$ such that $\theta_y \in [-1, 0]$ for all $y \in (m(X) - \tilde{\delta}_0, m(X)]$. Then the above discussion implies $\Lambda''_X(\theta_y)$ is a continuous function of y for all $y \in (m(X) - \tilde{\delta}_0, m(X)]$. Furthermore, for y in this range, $\Lambda'_X(\theta_y) < \infty$ and $\Lambda''_X(0) > 0$. This gives $\infty > \Lambda''_X(\theta_y) > 0$ and $c := \sup\left\{\Lambda''_X(\theta_y) : \ y \in [m(X) - \tilde{\delta}_0, m(X)]\right\}$ is finite (continuous function on a compact set).

Now, it can be checked that for $y \in [m(X) - \tilde{\delta}_0, m(X)]$, $I'(y) = \theta_y$ and $I''(y) = \frac{1}{\Lambda''(\theta_y)} \geq c^{-1}$. Moreover, $I(m(X) - \delta) = I(m(X)) + I'(m(X))\delta + \delta^2/2I''(\tilde{x})$, for $\tilde{x} \in [m(X) - \delta, m(X)]$. This gives

$$I(m(X) - \delta) \geq \frac{\delta^2}{2c}, \quad \text{for } \delta \leq \delta_0.$$

■

B.3. Bounding deviations of sub-optimal arms

Recall that for $T \geq K + 1$, N denotes the random number of batches played by the algorithm till time T, B_j denotes the random number of samples allocated within the j^{th} batch, and T_j denotes the

time of beginning of the j^{th} batch. Furthermore, recall

$$E_{1j} = \mathbb{1}\left(\mathrm{KL}_{\inf}\left(\hat{\mu}_a(T_j), m\right) \leq \frac{g_a(T_j)}{N_a(T_j)},\ \mathrm{KL}_{\inf}\left(\hat{\mu}_a(T_j), m\right) > \mathrm{KL}_{\inf}(\mu_a, m) - \delta,\ A_{T_j} = a\right).$$

This corresponds to the event that sufficient samples have not been allocated to the sub-optimal arm a, and contributes to the regret of the algorithm. Clearly,

$$\sum_{j=1}^{N} B_j E_{1j} \leq \sum_{j=1}^{N} B_j \mathbb{1}\left(N_a(T_j) \leq \frac{g_a(T_j)}{\mathrm{KL}_{\inf}\left(\mu_a, m\right) - \delta}, A_{T_j} = a\right).$$

Next, recall that for sub-optimal arm a,

$$E_{2j} = \mathbb{1}\left(\mathrm{KL}_{\inf}(\hat{\mu}_a(T_j), m) \leq \mathrm{KL}_{\inf}(\mu_a, m) - \delta,\ A_{T_j} = a\right).$$

Let $N_{B,a}$ denote the random number of batches allocated to arm a till time T and B_t denote the size of the batch begining at time t.

Lemma 14 *For $T \geq K+1$, $\tilde{\eta} \geq 0$, $\delta > 0$,*

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

Proof Since $\log(t) + 2\log\log(t)$ is a monotonically increasing function, using the form of $g_a(.)$,

$$\sum_{j=1}^{N} B_j E_{1j} \leq \sum_{j=1}^{N} B_j \mathbb{1}\left(N_a(T_j) \leq \frac{\log(T) + 2\log\log(T) + 2\log(1+N_a(T_j)) + 1}{\mathrm{KL}_{\inf}\left(\mu_a, m\right) - \delta},\ A_{T_j} = a\right).$$

Clearly, $z^* := \sup\left\{z \in \mathbb{N} : zd \leq \log(T) + 2\log\log(T) + 2\log(1+z) + 1\right\}$ is at most $\tilde{z}$, which equals

$$\frac{\log(T) + 2\log\log(T)}{d} + \left(1 + \frac{2}{d}\right)\log\left(1 + \frac{\log(T) + 2\log\log(T)}{d}\right) + \frac{10}{d} + O\left(\log\log\log(T)\right). \tag{17}$$

Thus, setting $d = \mathrm{KL}_{\inf}(\mu_a, m) - \delta$, we get that

$$\sum_{j=1}^{N} B_j E_{1j} \leq B_N \mathbb{1}\left(N_a(T_N) \leq \frac{\log(T) + 2\log\log(T) + 2\log(1+N_a(T_N)) + 1}{\mathrm{KL}_{\inf}\left(\mu_a, m\right) - \delta},\ A_{T_j} = a\right) + \tilde{z}.$$

Clearly, when the indicator above is 1, then B_N is at most $(\tilde{\eta}\tilde{z} + 1)$. Thus,

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

where the lower order terms in the above expression are the $o(\log(T))$ terms in (17), with $d = \mathrm{KL}_{\inf}(\mu_a, m) - \delta$. ■

Lemma 15 *For $T > K$, for sub-optimal arm a,*

$$\mathbb{E}\left(\sum_{j=1}^{N} B_j E_{2j}\right) \leq \begin{cases} \frac{1+\tilde{\eta}}{c_\mu \delta^2}\left(\frac{1}{\log(1+\tilde{\eta})} + \frac{1}{e}\right), \textit{ for } \tilde{\eta} > 0 \\ \frac{1}{c_\mu \delta^2} + 1, \textit{ otherwise.} \end{cases}$$

Proof Recall that

$$\mathbb{E}\left(\sum_{j=1}^{N} B_j E_{2j}\right) = \mathbb{E}\left(\sum_{k=2}^{N_{B,a}} B_{T_a^k} \mathbb{1}\left(\mathrm{KL}_{\inf}(\hat{\mu}_a(T_a^k), m) \leq \mathrm{KL}_{\inf}(\mu_a, m) - \delta\right)\right),$$

where T_a^k denotes the random time of beginning of the batch when arm a won for the k^{th} time.

Let us first consider the case when $\tilde{\eta} > 0$. In this case, $N_{B,a}$ is at most $\frac{\log(T)}{\log(1+\tilde{\eta})}$ and $B_{T_a^k}$ is at most $\tilde{\eta} N_a(T_a^k) + 1$, which in turn is at most $(1+\tilde{\eta})^{k-1}$. Thus, the required expectation is at most

$$\sum_{k=2}^{\frac{\log(T)}{\log(1+\tilde{\eta})}+1} (1+\tilde{\eta})^{k-1} \mathbb{P}\left(\mathrm{KL}_{\inf}(\hat{\mu}_a(T_a^k), m) \leq \mathrm{KL}_{\inf}(\mu_a, m) - \delta\right).$$

Clearly, $N_a(T_a^k)$ is deterministic. Lemma 6 bounds the probability in the above expression by $e^{-N_a(T_a^k) c_\mu \delta^2}$. Lemma 16 then bounds the summation, giving the desired bound in this case.

When $\tilde{\eta} = 0$, $N_{B,a} \leq T$, $B_{T_a^k} = 1$ and $T_a^k \geq K + k - 1$. The required average is bounded by

$$\sum_{k=2}^{T} e^{-k c_\mu \delta^2} \leq \frac{1}{c_\mu \delta^2} + 1,$$

giving the desired bound. ■

Lemma 16 *For $\delta > 0$, there exists a constant $\tilde{c}_\mu(\delta)$ (independent of T) such that*

$$\sum_{k=1}^{\frac{\log T}{\log(1+\tilde{\eta})}+1} (1+\tilde{\eta})^{k-1} e^{-N_a(T_a^k) c_\mu \delta^2} \leq \tilde{c}_\mu(\delta).$$

Proof Recall that $(1+\tilde{\eta})^{k-2} \leq N_a(T_a^k)$. The required summation is bounded by

$$(1+\tilde{\eta}) \sum_{k=1}^{\frac{\log T}{\log(1+\tilde{\eta})}+1} (1+\tilde{\eta})^{k-2} e^{-(1+\tilde{\eta})^{k-2} c_\mu \delta^2},$$

which is further bounded by

$$(1+\tilde{\eta}) \int_1^{\infty} (1+\tilde{\eta})^{k-2} e^{-(1+\tilde{\eta})^{k-2} c_\mu \delta^2} dk + \frac{1+\tilde{\eta}}{c_\mu \delta^2 e},$$

where, $\frac{1}{c_\mu \delta^2 e}$ is the maximum value of the function being summed. To see the above bound, we upper bound the required summation by sum of the integral of the function from 1 to ∞ and the maximum value of the function. The above can then be shown to equal

$$\tilde{c}_\mu(\delta) := \frac{1+\tilde{\eta}}{c_\mu \delta^2}\left(\frac{e^{-c_\mu \delta^2/(1+\tilde{\eta})}}{\log(1+\tilde{\eta})} + \frac{1}{e}\right),$$

which is the required upper bound. ■

B.4. Bounding the deviations of optimal arm

Recall that, for $k \geq 2$, T_a^k denotes the random time of the beginning of the batch when the a^{th} arm won for the k^{th} time. In particular, arm a has been sampled for $k-1$ batches till this time. Thus, T_a^k is at least $K - 1 + 1 + \tilde{\eta} + \cdots + \tilde{\eta}(1+\tilde{\eta})^{k-3} = K - 1 + (1+\tilde{\eta})^{k-2}$. Using this bound on T_a^k, the following lemma follows directly from Proposition 5.

Lemma 17 *For $k \geq 2$, $g_1(t) = \log(t) + 2\log\log(t) + 2\log(1 + N_1(t)) + 1$,*

$$\mathbb{P}\left(N_1(T_a^k)\,\mathrm{KL}_{\inf}\left(\hat{\mu}_1(T_a^k), m(\mu_1)\right) \geq g_1(T_a^k)\right) \leq (1+\tilde{\eta})^{-k+2}\left(\log\left(K - 1 + (1+\tilde{\eta})^{k-2}\right)\right)^{-2}.$$

Now, recall that for $m = m(\mu_1)$,

$$D_N := \sum_{j=K+1}^{N} B_j \mathbb{1}\left(U_1(T_j) \leq m,\ A_{T_j} = a\right).$$

Lemma 18 *For $T > K$,*

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

Proof Recall that $N_{B,a}$ denotes the number of batches allocated to arm a in time T, T_a^k denotes the time of beginning of batch when arm a won for the k^{th} time, and $m = m(\mu_1)$. Then D_N can be re-written as

$$\sum_{k=2}^{N_{B,a}(T)} B_{T_a^k} \mathbb{1}\left(U_1(T_a^k) \leq m\right).$$

Recall that for any t, the event $\{U_1(t) \leq m\}$ is same as $\{N_1(t)\,\mathrm{KL}_{\inf}(\hat{\mu}_1(t), m) \geq g_1(t)\}$, giving

$$D_N = \sum_{k=2}^{N_{B,a}} B_{T_a^k} \mathbb{1}\left(N_1(T_a^k)\,\mathrm{KL}_{\inf}\left(\hat{\mu}_1(T_a^k), m\right) \geq g_1(T_a^k)\right).$$

Let us first consider the case when $\tilde{\eta} > 0$. In this case, $N_{B,a}$ is at most $\frac{\log(T)}{\log(1+\tilde{\eta})}$ and $B_{T_a^k}$ is at most $\tilde{\eta} N_a(T_a^k) + 1$, which in turn is at most $(1+\tilde{\eta})^{k-1}$. Thus, the required expectation is at most

$$\mathbb{E}\left(D_N\right) \leq \sum_{k=2}^{\frac{\log(T)}{\log(1+\tilde{\eta})}+1} (1+\tilde{\eta})^{k-1}\,\mathbb{P}\left(N_1(T_a^k)\,\mathrm{KL}_{\inf}\left(\hat{\mu}_1(T_a^k), m\right) \geq g_1(T_a^k)\right).$$

For $g_1(t) = \log(t) + 2\log\log(t) + 2\log(1 + N_a(t)) + 1$, Lemma 17 bounds the probability in the expression in the r.h.s. above. Summing over $k \in \{2, \ldots \log(T)/\log(1+\tilde{\eta})\}$, we get

$$\mathbb{E}\left(D_N\right) \leq (1+\tilde{\eta})\left(\frac{1}{(\log K)^2} + \frac{\pi^2}{6(\log(1+\tilde{\eta}))^2}\right).$$

Now, let $\tilde{\eta} = 0$. In this case, $N_{B,a} \leq T$, $B_{T_a^k} = 1$, $T_a^k \geq K + k - 1$. Using these, together with $g_1(T_a^k) \geq \log(K+k-1) + 2\log\log(K+k-1) + 2\log(1 + N_1(T_a^k)) + 1$, we get

$$\mathbb{E}\left(D_N\right) \leq \sum_{k=2}^{T} \mathbb{P}\left(N_1(T_a^k)\,\mathrm{KL}_{\inf}(\hat{\mu}_1(T_a^k), m) \geq g_1(T_a^k)\right) \leq \sum_{k=2}^{T} (K+k-1)^{-1}\left(\log(K+k-1)\right)^{-2},$$

which is bounded by the constant in the statement. ■

Appendix C. Comparison with Robust-UCB of Bubeck et al. (2013)

Consider the following optimization problem that corresponds to our index:

$$\max_{\kappa \in \mathcal{L}} \ \mathbb{E}_\kappa\left(X\right) \quad \text{s.t.} \quad n\,\mathrm{KL}(\hat{\eta}(n), \kappa) \leq C. \tag{18}$$

Recall that the Donsker-Varadhan variational representation for KL-divergence expresses the KL-divergence between any two probability measures P, Q, defined on a common space Ω, as

$$\mathrm{KL}(P,Q) = \sup_g \left\{\mathbb{E}_P\left(g(X)\right) - \log \mathbb{E}_Q\left(e^{g(X)}\right)\right\},$$

where the supremum is taken over all measurable functions $g : \Omega \to \Re$ such that $\mathbb{E}_Q\left(e^{g(X)}\right)$ is well-defined. Using this to lower-bound KL with a specific choice of g in the r.h.s., relaxes the constraint in index-optimization problem, giving the following upper bound on our index:

$$\max_{\kappa \in \mathcal{L}} \ \mathbb{E}_\kappa\left(X\right) \quad \text{s.t.} \quad n\mathbb{E}_{\hat{\eta}(n)}\left(g(X)\right) - n\log \mathbb{E}_\kappa\left(e^{g(X)}\right) \leq C.$$

For a sequence of thresholds u_n (to be specified later), and $\theta > 0$, define a function $g_n(X) = -\theta X \mathbb{1}\left(|X| \leq u_n\right)$. Substituting g_n for g in the above, and adding $n\theta\mathbb{E}_\kappa\left(X\right)$ on both the sides, we get that our index is bounded from above be $\max\ \mathbb{E}_\kappa\left(X\right)$ such that $\kappa \in \mathcal{L}$ and

$$\theta\sum_{i=1}^{n}\left(\mathbb{E}_\kappa\left(X\right) - X_i\mathbb{1}\left(|X_i| \leq u_n\right)\right) - n\log \mathbb{E}_\kappa\left(e^{-\theta X \mathbb{1}(|X| \leq u_n)}\right) \leq C + n\theta\mathbb{E}_\kappa\left(X\right). \tag{19}$$

Let $Y_n := X\mathbb{1}\left(|X| \leq u_n\right)$ and $m_n := \mathbb{E}_\kappa\left(X\mathbb{1}\left(|X| \leq u_n\right)\right)$. Then, $\mathbb{E}_\kappa\left(\theta^2 Y_n^2\right) \leq \theta^2 B u_n^{1-\epsilon}$, and

$$\mathbb{E}_\kappa\left(e^{-\theta X \mathbb{1}(|X| \leq u_n)}\right) \leq 1 - \theta m_n + \sum_{j=2}^{\infty}\frac{\mathbb{E}_\kappa\left(|\theta Y_n|^j\right)}{j!} \leq 1 - \theta m_i + \frac{B}{u_n^{1+\epsilon}}\sum_{j=2}^{\infty}\frac{(\theta u_n)^j}{j!}. \tag{20}$$

Thus, we have $\mathbb{E}_\kappa\left(e^{-\theta X \mathbb{1}(|X|\le u_n)}\right) \le 1 - \theta m_n + \frac{B}{u_n^{1+\epsilon}}\left(e^{\theta u_n} - \theta u_n - 1\right)$. Using $1 + x \le e^x$ and (20) in (19), we get that the optimal value of the following optimization problem is an upper bound on our index: $\max\ \mathbb{E}_\kappa(X)$ subject to $\kappa \in \mathcal{L}$ and

$$\theta \sum_{i=1}^{n} \left(\mathbb{E}_\kappa(X) - X_i \mathbb{1}(|X_i| \le u_n)\right) \le C + n\left(\theta \mathbb{E}_\kappa(X) - \theta m_n + \frac{B}{u_n^{1+\epsilon}}\left(e^{\theta u_n} - \theta u_n - 1\right)\right).$$

Since $\kappa \in \mathcal{L}$, $\mathbb{E}_\kappa(X \mathbb{1}(|X| \ge u_n))$ is at most $\frac{B}{(u_n)^\epsilon}$, and the constraint above can be re-arranged, and further relaxed to

$$\frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_\kappa(X) - X_i \mathbb{1}(|X_i| \le u_n)\right) \le \frac{B}{u_n^\epsilon} + \frac{1}{\theta}\left(\frac{C}{n} + \frac{B}{u_n^{1+\epsilon}}\left(e^{\theta u_n} - \theta u_n - 1\right)\right).$$

Choosing $u_n = \left(\frac{Bn}{\log \delta^{-1}}\right)^{\frac{1}{1+\epsilon}}$ and $\theta = \frac{C u_n^\epsilon}{nB}$, the above constraint is $\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_\kappa(X) - X_i \mathbb{1}(|X_i| u_n)\right)$ less than

$$\sum_{i=1}^{n} B\left(\frac{Bn}{\log \delta^{-1}}\right)^{-\frac{\epsilon}{1+\epsilon}} + (nB)^{\frac{1}{1+\epsilon}} \frac{\log \delta^{-1}}{C}\left(\log \delta^{-1}\right)^{\frac{\epsilon}{1+\epsilon}}\left(e^{\frac{C}{\log \delta^{-1}}} - 1\right).$$

Setting $\hat{\mu}_T(n) = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}(|X_i| \le u_n)$ and re-arranging, we get the following bound on the index:

$$\hat{\mu}_T(n) + B^{\frac{1}{1+\epsilon}}\left(\frac{\log \delta^{-1}}{n}\right)^{\frac{\epsilon}{1+\epsilon}}\left(1 + \left(e^{\frac{C}{\log \delta^{-1}}} - 1\right)\frac{\log \delta^{-1}}{C}\right).$$

Let us conclude this section with a remark about a phase transition in $\epsilon = 1$. The bound in the previous display, which is valid for $\epsilon \le 1$, is of order $n^{\frac{-\epsilon}{1+\epsilon}}$. One may wonder what happens for $\epsilon > 1$, but one immediately concludes that the bound cannot hold beyond $\epsilon > 1$. This is due to the presence of distributions in $\mathcal{L}$ for which the central limit theorem holds (for example standard Rademacher when $B \ge 1$), which force the width to be at least $n^{-\frac{1}{2}}$. The $\mathrm{KL}_{\inf}$-based approach (18) does allow $\epsilon > 1$. No phase transition is present in the formulas, though a phase transition may be present in the set of constraints active at the optimum. For example, in Lemma 19 below, ϵ determines whether the lower bound on λ_1 is active (which corresponds to whether the support of κ includes one additional point beyond the support of η) or not.

Appendix D. Computing the index

Let $\mathcal{L} \subseteq \mathcal{P}(\Re)$ be the set of distributions on $\Re$ with $(1+\epsilon)^{\text{th}}$ moment bounded by B. For $\eta \in \mathcal{P}(\Re)$ itself possibly outside $\mathcal{L}$, we define the upper index at threshold C by (5), which we re-state

$$U_\eta \ := \ \max_\kappa\ m(\kappa) \qquad \text{s.t.} \qquad \kappa \in \mathcal{L} \quad \text{and} \quad \mathrm{KL}(\eta, \kappa) \le C.$$

We now give a characterization of the above optimisation problem. Note that the variation with constraint $n\,\mathrm{KL}(\eta, \kappa) \le C$ follows from the below result after dividing C by n, the number of samples.

Lemma 19 (Dual formulation of index U_η)

$$U_\eta \;=\; \min_{\lambda_1,\lambda_2} \lambda_1 + \lambda_2 B - e^{-C+\int \eta(x)\ln\left(\lambda_1 - x + \lambda_2|x|^{1+\epsilon}\right)\mathrm{d}x} \quad \textit{s.t.} \quad \lambda_1 \geq \frac{\epsilon\lambda_2^{-\frac{1}{\epsilon}}}{(1+\epsilon)^{1+\frac{1}{\epsilon}}} \quad \textit{and} \quad \lambda_2 \geq 0.$$

Proof The proof is a diligent application of convex duality. Introducing Lagrange multipliers λ_1, λ_2 and λ_3 for the normalisation, moment and KL-ball constraints respectively, we find that the objective equals

$$\min_{\lambda_1,\lambda_2\geq 0,\lambda_3\geq 0} \max_{\kappa\geq 0} \mathbb{E}_\kappa[X] + \lambda_1\left(1 - \mathbb{E}_\kappa\left[1\right]\right) + \lambda_2\left(B - \mathbb{E}_\kappa\left[|X|^{1+\epsilon}\right]\right) + \lambda_3\left(C - \mathrm{KL}(\eta,\kappa)\right).$$

The solution for κ is

$$\kappa(x) \;=\; \frac{\lambda_3\eta(x)}{\lambda_1 - x + \lambda_2|x|^{1+\epsilon}},$$

where we inherit the restriction $(\lambda_1,\lambda_2) \in \mathcal{R} := \left\{(\lambda_1,\lambda_2) \in \mathbb{R}^2 \middle| \forall x \in \mathbb{R} : \lambda_1 - x + \lambda_2|x|^{1+\epsilon} \geq 0\right\}$ (which in particular implies that $\lambda_1 \geq 0$, and κ has at most one additional support point compared to η, which must then be at $(\lambda_2(1+\epsilon))^{-1/\epsilon}$). Plugging this in, we find

$$\min_{(\lambda_1,\lambda_2)\in\mathcal{R},\lambda_3\geq 0} -\lambda_3 + \lambda_3\int \eta(x)\ln\left(\frac{\lambda_3}{\lambda_1 - x + \lambda_2|x|^{1+\epsilon}}\right)\mathrm{d}x + \lambda_1 + \lambda_2 B + \lambda_3 C.$$

Optimising for λ_3 gives

$$\lambda_3 \;=\; e^{\int \eta(x)\ln\left(\lambda_1 - x + \lambda_2|x|^{1+\epsilon}\right)\mathrm{d}x - C},$$

and plugging this in gives the claim. ■

The upshot of this result is that our $\mathrm{KL}_{\inf}$-based indices can be computed with convex optimisation tools. The optimisation variable has dimension 2, making standard convex optimisation including e.g. the ellipsoid method practical. Note that the optimisation region for (λ_1,λ_2) is unbounded, which may be addressed by successively enlarging the starting ellipsoid. When applying this to an empirical distribution η supported on n points, the number of terms in the objective (and hence the run time) scales linearly with n and also with the number of bits of precision required.

Appendix E. Finite time bound for bounded support distributions: Proof of Proposition 4

In this section, we establish the conjectured optimality of the empirical KL-UCB algorithm of Cappé et al. (2013b) and give the first optimal finite-time regret bound for bounded-support arm distributions. In this setting, $\mathcal{L} = \mathcal{P}([0,1])$, and for $\eta \in \mathcal{P}(\Re)$, $\mathrm{KL}^L_{\inf}(\eta,x)$ is defined to be $\inf \mathrm{KL}(\eta,\kappa) : \kappa \in \mathcal{L},$ and $m(\kappa) \geq x$. Honda and Takemura (2010) develop alternate representations for the $\mathrm{KL}^{\mathcal{L}}_{\inf}$ in this setting. They show that

$$\mathrm{KL}^{\mathcal{L}}_{\inf}(\mu_a,x) := \max_{\lambda\in\left[0,\frac{1}{1-x}\right]} \mathbb{E}_{\mu_a}\left(\log\left(1-(X-x)\lambda\right)\right) \qquad \text{for } x \in [0,1].$$

$\mathrm{KL}_{\inf}$-UCB, with $\mathcal{L} = \mathcal{P}([0,1])$, $\mathrm{KL}_{\inf}^{\mathcal{L}}$ defined above, and $g(t) = \log(t) + 2\log(\log(t))$, recovers the empirical KL-UCB algorithm of Cappé et al. (2013b). In particular, the index for arm a, denoted as $U_a(t)$, is given by

$$U_a(t) = \max\left\{x : \mathrm{KL}_{\inf}^{\mathcal{L}}(\hat{\mu}_a(t), x) \leq \frac{g(t)}{N_a(t)}\right\}.$$

We highlight the key steps in the proof of the finite time bound, below.

As earlier, we bound the average number of pulls of a sub-optimal arm, a. For simplicity of notation, let us assume that arm 1 is the arm with maximum mean. Let $\epsilon_1 \in (0, m(\mu_1))$, and $\tilde{m} = m(\mu_1) - \epsilon_1$. Then for arm $a \neq 1$, using the definition of the index, we have

$$N_a(T) = 1 + \sum_{t=K+1}^{T} \mathbb{1}\left(A_t = a\right) = D_T + E_T + 1,$$

where A_t denotes the arm selected at time t, and

$$D_T := \sum_{t=K+1}^{T} \mathbb{1}\left(\mathrm{KL}_{\inf}^{\mathcal{L}}\left(\hat{\mu}_1(t), \tilde{m}\right) \geq \frac{g(t)}{N_1(t)} \text{ and } A_t = a\right),$$

and

$$E_T := \sum_{t=K+1}^{T} \mathbb{1}\left(\mathrm{KL}_{\inf}^{\mathcal{L}}\left(\hat{\mu}_a(t), \tilde{m}\right) < \frac{g(t)}{N_a(t)} \text{ and } A_t = a\right).$$

Using Cappé et al. (2013a, Section B.2, (26)) and the bounds therein, and that $\epsilon_1 < 1$,

$$\mathbb{E}\left(D_T\right) \leq \left(\sum_{t=K+1}^{T} e^{-g(t)}\right)\left(3e + 2 + \frac{4}{\epsilon_1^2} + \frac{8e}{\epsilon_1^4}\right) \leq \frac{36}{\epsilon_1^4}\left(\sum_{t=K+1}^{T} e^{-g(t)}\right).$$

For $t \geq 2$ and $K \geq 2$ we have

$$\mathbb{E}\left(D_T\right) \leq \frac{36}{\epsilon_1^4}\sum_{t=K+1}^{T}\frac{1}{t\log t} \leq \frac{36}{\epsilon_1^4}\left(\frac{1}{2\log 2} + \int_2^T \frac{1}{t\log t}\,dt\right) \leq \frac{36}{\epsilon_1^4}\left(2 + \log\log T\right). \tag{21}$$

Moreover, $\mathbb{E}\left(E_T\right)$ is bounded by

$$\sum_{t=K+1}^{T} \mathbb{P}\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\hat{\mu}_a(t), \tilde{m}) \leq \frac{g(t)}{N_a(t)};\ \ \mathrm{KL}_{\inf}^{\mathcal{L}}(\hat{\mu}_a(t), \tilde{m}) \geq \mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, \tilde{m}) - \delta;\ \ A_t = a\right)$$
$$+ \sum_{t=K+1}^{T} \mathbb{P}\left(A_t = a;\ \ \mathrm{KL}_{\inf}^{\mathcal{L}}(\hat{\mu}_a(t), \tilde{m}) \leq \mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, \tilde{m}) - \delta\right),$$

which is further bounded by

$$\sum_{t=K+1}^{T} \mathbb{P}\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, \tilde{m}) - \delta \leq \frac{g(t)}{N_a(t)};\ \ A_t = a\right) + \sum_{s=1}^{T} \mathbb{P}\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\hat{\mu}_{a,s}, \tilde{m}) \leq \mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, \tilde{m}) - \delta\right),$$

where $\hat{\mu}_{a,s}$ denotes the empirical distribution for arm a, corresponding to s samples from that arm. Clearly, the first term in the above summation is at most

$$g(T)\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, \tilde{m}) - \delta\right)^{-1} + 1,$$

while the second term is bounded using Honda and Takemura (2015, Theorem 12) by

$$\sum_{s=1}^{T} e^{-sc(\delta,\epsilon_1)} \leq \left(1 - e^{-c(\delta,\epsilon_1)}\right)^{-1}, \quad \text{where} \quad c(\delta,\epsilon_1) = \frac{\delta^2}{2\left(c_0 + \frac{1-m(\mu_a)}{1-\tilde{m}}\right)}, \tag{22}$$

with the condition that $\delta \leq \frac{1}{2}\left(c_0 + \frac{1-m(\mu_a)}{1-\tilde{m}}\right)$, and $c_0 \geq 2.2$. Thus, we have

$$\mathbb{E}\left(E_T\right) \leq g(T)\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, \tilde{m}) - \delta\right)^{-1} + 1 + \frac{1}{1 - e^{-c(\delta,\epsilon_1)}},$$

where $c(\delta,\epsilon_1)$ is specified above. Using Cappé et al. (2013a, Lemma 4),

$$\mathbb{E}\left(E_T\right) \leq g(T)\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, m(\mu_1)) - \frac{\epsilon_1}{1-m(\mu_1)} - \delta\right)^{-1} + 1 + \left(1 - e^{-c(\delta,\epsilon_1)}\right)^{-1}, \tag{23}$$

Using (21) and (23) above, average number of pulls of a sub-optimal arm is bounded by

$$g(T)\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, m(\mu_1)) - \frac{\epsilon_1}{1-m(\mu_1)} - \delta\right)^{-1} + 1 + \left(1 - e^{-c(\delta,\epsilon_1)}\right)^{-1} + \frac{36}{\epsilon_1^4}\left(2 + \log\log T\right),$$

with $c(\delta,\epsilon_1)$ specified in (22). This bound can be optimized over ϵ_1 and δ under the conditions that $\epsilon_1 < m(\mu_1)$ and $\delta \leq \frac{1}{2}\left(c_0 + \frac{1-m(\mu_a)}{1-m(\mu_1)+\epsilon_1}\right)$. Choosing

$$\delta^3 = 8\left(c_0 + \frac{1-m(\mu_a)}{1-m(\mu_1)+\epsilon_1}\right)\left(\frac{\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, m(\mu_1))\right)^2}{g(T)}\right)$$

and

$$\epsilon_1^5 = \frac{\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, m(\mu_1))\right)^2}{g(T)}\left(2 + \log\log T\right),$$

the number of times a sub-optimal arm is pulled is bounded by

$$(\log T + \log\log T)\left(\mathrm{KL}_{\inf}^{\mathcal{L}}(\mu_a, m(\mu_1)) - O\left(\left(\frac{\log\log T}{\log T}\right)\right)^{\frac{1}{5}}\right)^{-1} + O\left((\log T)^{\frac{4}{5}}\left(\log\log(T)\right)^{\frac{1}{5}}\right).$$

Appendix F. Proving Theorem 3

In this section, we establish the theoretical guarantees for $\mathrm{KL}_{\inf}$-UCB2 algorithm, which is a perturbed version of $\mathrm{KL}_{\inf}$-UCB. Recall that for $\epsilon_1 > 0$, we define $\tilde{B} = B + \epsilon_1$, and let $\delta_t' = \log(1 + (\log\log(t))^{-1})$. Given $\mu \in \mathcal{L}^K$, for $\tilde{\eta} \geq 0$, $\mathrm{KL}_{\inf}$-UCB2 is precisely $\mathrm{KL}_{\inf}$-UCB(K,$\tilde{B}$,ϵ, $\tilde{\eta}$, $(1+\delta_t')^2\log(t)$). Proof of Theorem 3 follows exactly along the lines of Theorem 1. We highlight only the differences here.

As earlier, we analyse the events leading to selection of a sub-optimal arm by the algorithm. For $\epsilon_2 > 0$, in this section, let $m = m(\mu_1) - \epsilon_2$. The event that at the beginning of j^{th} batch, sub-optimal arm, a, has the maximum index, i.e., $\left\{A_{T_j} = a\right\}$ for $a \neq 1$, equals

$$\left\{U_1(T_j) \leq m \quad \text{and} \quad A_{T_j} = a\right\} \bigcup \left\{U_a(T_j) > m \quad \text{and} \quad A_{T_j} = a\right\}. \tag{24}$$

Let N denote the random number of batches till time T. As earlier, the random variable of interest is

$$\mathbb{E}\left(N_a(T)\right) = 1 + \mathbb{E}\left(\sum_{j=K+1}^{N} B_j \mathbb{1}\left(A_{T_j} = a\right)\right) \;=\; 1 + \mathbb{E}\left(D_N\right) + \mathbb{E}\left(E_N\right),$$

where, using the division from (24), we define

$$D_N := \sum_{j=K+1}^{N} B_j \mathbb{1}\left(U_1(T_j) \leq m,\; A_{T_j} = a\right), \text{ and } E_N := \sum_{j=K+1}^{N} B_j \mathbb{1}\left(U_a(T_j) > m,\; A_{T_j} = a\right).$$

Proof for controlling the deviations of the sub-optimal arm, i.e., $\mathbb{E}\left(E_N\right)$ above, follows exactly as earlier, with $\mathrm{KL}_{\inf}(\mu_a, m(\mu_1))$ replaced by $\mathrm{KL}^{\epsilon_1}_{\inf}(\mu_a, m(\mu_1) - \epsilon_2)$. Thus for any $\delta > 0$, we will have

$$\mathbb{E}\left(E_N\right) \leq \frac{(1+\tilde{\eta})(1+\delta_t')^2 \log(t)}{\mathrm{KL}^{\epsilon_1}_{\inf}(\mu_a, m(\mu_1) - \epsilon_2) - \delta} + O(1),$$

where $O(1)$ terms involve constants that are functions of ϵ_1, ϵ_2, and δ.

Let T_a^k denote the time of beginning of the batch when arm a won for the k^{th} time. Let us now show that $\mathbb{E}\left(D_N\right) = O(1)$.

As earlier,

$$\mathbb{E}\left(D_N\right) \leq \sum_{k=2}^{\frac{\log(T)}{\log(1+\tilde{\eta})}+1} (1+\tilde{\eta})^{k-1}\, \mathbb{P}\left(N_1(T_a^k)\, \mathrm{KL}_{\inf}\left(\hat{\mu}_1(T_a^k), m\right) \geq g_1(T_a^k)\right). \tag{25}$$

Proposition 20 bounds the probability in the summand above. Lemma 21 argues that $\mathbb{E}\left(D_N\right) = o(\log T)$.

Thus,

$$\mathbb{E}\left(N_a(T)\right) \leq \frac{(1+\tilde{\eta})(1+\delta_t')^2 \log T}{\mathrm{KL}^{\epsilon_1}_{\inf}(\mu_a, m(\mu_1) - \epsilon_2) - \delta} + o(\log T).$$

Dividing by $\log(T)$ and taking limit as $T \to \infty$,

$$\limsup_{T\to\infty} \frac{\mathbb{E}\left(N_a(T)\right)}{\log(T)} \leq \frac{1+\tilde{\eta}}{\mathrm{KL}^{\epsilon_1}_{\inf}(\mu_a, m(\mu_1) - \epsilon_2) - \delta}.$$

Since $\epsilon_2 > 0$ and $\delta > 0$ are arbitrary constants, taking infimum over these, and using that $\mathrm{KL}^{\epsilon_1}_{\inf}(\mu_a, x)$ is a continuous function of x, we get

$$\limsup_{T\to\infty} \frac{\mathbb{E}\left(N_a(T)\right)}{\log(T)} \leq \frac{1+\tilde{\eta}}{\mathrm{KL}^{\epsilon_1}_{\inf}(\mu_a, m(\mu_1))}. \qquad \square$$

For $x \in \Re$, let $f(x) = |x|^{1+\epsilon}$, and for $c \geq 0$ define $f^{-1}(c) := c^{1/(1+\epsilon)}$. Let $T > t > 0$, $\gamma > 0$, $X \sim \mu_1$,

$$C_1 := C_2 \left(1 - e^{\frac{-\epsilon_2}{f^{-1}(\tilde{B}) - m}}\right)^{-1} \left(1 - e^{\frac{-\epsilon_1}{\tilde{B} - f(m)}}\right)^{-1},$$

$$C_T := \left(1 - (1+\gamma)^{-1-\delta_T'}\right) \text{ and } C_2 := \exp\left(\frac{\mathbb{E}\left(|X - \tilde{m}|\right)}{f^{-1}(\tilde{B}) - m} + \frac{\mathbb{E}\left(\left|\tilde{B} - f(X)\right|\right)}{\tilde{B} - f\left(m\right)}\right).$$

Proposition 20 *For $T > K$, $k > 0$, $\epsilon_1 > 0$, and $g(t) = (1 + \delta_t')^2 \log(t)$,*

$$\mathbb{P}\left(\mathrm{KL}_{\inf}^{\epsilon_1}(\hat{\mu}_1(T_a^k), m) \geq \frac{g(T_a^k)}{N_1(T_a^k)}\right) \leq \frac{C_1 \left(1 + \tilde{\eta}\right)^{-(k-1)(1+\delta_T')}}{\log\left(1 + \delta_T'\right)} \left(\frac{\log\left(1 + \tilde{\eta}\right)^{k-1}}{C_T} + \frac{\log\left(1 + \gamma\right)}{C_T^2}\right).$$

In particular, introducing the perturbations allows us to get rid of the additional $2\log(1 + N_1(t))$ cost in the threshold, which resulted from the presence of this term in the bound in Proposition 5. Proposition 20 is proved in Section F.1 below.

Lemma 21 *For $T > 0$, $\epsilon_1 > 0$ and $g(t) := \left(1 + \log\left(1 + \frac{1}{\log\log t}\right)\right)^2 \log(t)$,*

$$\sum_{k=1}^{\frac{\log(T)}{\log(1+\tilde{\eta})}+1} \left(1 + \tilde{\eta}\right)^k \mathbb{P}\left(N_1(T_a^k)\, \mathrm{KL}_{\inf}^{\epsilon_1}\left(\hat{\mu}_1(T_a^k), m\right) \geq g(T_a^k)\right) = o(\log\left(T\right)).$$

Proof Using Proposition 20 to bound the probability in the summation, the required expression can be bounded by

$$\left(1 + \tilde{\eta}\right) \sum_{k=1}^{\frac{\log T}{\log(1+\tilde{\eta})}+1} \left(\frac{C_1}{C_T} \frac{\log\left(1 + \tilde{\eta}\right)}{\log\left(1 + \delta_T'\right)} \frac{k-1}{\left(1 + \tilde{\eta}\right)^{(k-1)\delta_T'}} + \frac{C_1 \log\left(1 + \gamma\right)}{C_T^2 \log\left(1 + \delta_T'\right)} \frac{1}{\left(1 + \tilde{\eta}\right)^{(k-1)\delta_T'}}\right),$$

where C_1, δ_T' and C_T are constants independent of k, such that C_T converges to a constant and δ_T' converges to 0, as $T \to \infty$ (see Proposition 20). It is then easy to see that

$$\left(\log\left(1 + \delta_T'\right)\left(1 - \frac{1}{\left(1 + \tilde{\eta}\right)^{\delta_T'}}\right)^2\right)^{-1} = o(\log T),$$

where $\delta_T' = \log\left(1 + \frac{1}{\log\log T}\right)$, and hence the required summation is $o(\log T)$. ■

F.1. Towards proving Proposition 20

In this section, we prove the concentration inequality in Proposition 20. The proof involves the use of peeling arguments, the dual formulation for $\mathrm{KL}_{\inf}^{\epsilon_1}$, ϵ-nets and careful use of martingales, and may also be of independent interest. To facilitate the proof, we need some notation and results that will be used later, which we prove first.

In this section, for $x \in \Re$, we define $f(x) := |x|^{1+\epsilon}$, and for $c \geq 0$ $f^{\text{-}1}(c) := c^{1/(1+\epsilon)}$. For $T > K$ and $K < t \leq T$, define $\delta'_t := \log\left(1 + \frac{1}{\log\log t}\right)$. For $\gamma > 0$, let $J_M = \frac{\log\left(T/(1+\tilde{\eta})^{k-1}\right)}{\log(1+\gamma)}$. For $j \in \{0, 1, \cdots, J_M\}$, define the event

$$D_j = \left\{(1+\tilde{\eta})^{k-1}(1+\gamma)^j \leq T_a^k \leq (1+\tilde{\eta})^{k-1}(1+\gamma)^{j+1}\right\},$$

and for $\eta > 0$, and $i \geq 0$, define

$$C_i(t) = \left\{(1+\eta)^i \leq N_1(t) \leq (1+\eta)^{i+1}\right\}.$$

Furthermore, for $\epsilon_2 > 0$ recall that $m = m(\mu_1) - \epsilon_2$.

Lemma 22 *For $\gamma > 0, \eta > 0$, $i \in \mathbb{N}, \tilde{\eta} \geq 0$, $k \geq 1$, $j \in [J_M]$,*

$$\mathbb{P}\left(N_1(T_a^k)\,\mathrm{KL}_{\inf}^{\epsilon_1}(\hat{\mu}_1(T_a^k), m) \geq g(T_a^k), D_j, C_i(T_a^k)\right) \leq C_1 e^{-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)}},$$

where

$$C_1 = \frac{C_2}{1 - e^{-\epsilon_2\left(f^{-1}(\tilde{B})-m\right)^{\text{-}1}}}\left(1 - e^{\frac{-\epsilon_1}{\tilde{B}-f(m)}}\right)^{-1} \quad \textit{and} \quad C_2 = e^{\frac{\mathbb{E}_{\mu_1}(|X-m|)}{f^{-1}(\tilde{B})-m}} e^{\frac{\mathbb{E}_{\mu_1}\left(|\tilde{B}-f(X)|\right)}{\tilde{B}-f(m)}}.$$

Proof On the set $C_i(T_a^k)$ (denoted as C_i in this proof), $N_1(T_a^k) \leq (1+\eta)^{i+1}$. Using this, the required probability is bounded from above by

$$\mathbb{P}\left(\mathrm{KL}_{\inf}^{\epsilon_1}(\hat{\mu}_1(T_a^k), m) \geq \frac{g(T_a^k)}{(1+\eta)^{i+1}}, D_j, C_i\right). \tag{26}$$

Using the dual formulation for $\mathrm{KL}_{\inf}^{\epsilon_1}$ from Lemma 7, above is bounded by

$$\mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\mathcal{R}(m,\tilde{B})} \sum_{l=1}^{N_1(T_a^k)} \log\left(1 - (X_l - m)\lambda_1 - (\tilde{B} - f(X_l))\lambda_2\right) \geq \frac{N_1(T_a^k) g(T_a^k)}{(1+\eta)^{i+1}}, D_j, C_i\right),$$

where $\mathcal{R}(m, \tilde{B})$ is a subset of $\Re^2$ similar to that defined in the Lemma 7, such that the argument of $\log$ in the expression above is always non-negative. Consider a (δ_1, δ_2)-net over the rectangle

$$\left[0, \frac{1}{f^{-1}(\tilde{B}) - m}\right] \times \left[0, \frac{1}{\tilde{B} - f(m)}\right],$$

which contains the region $\mathcal{R}(m, \tilde{B})$ (see Lemma 8). Let $\mathcal{G}_{l_1,l_2}$ denote a grid in the constructed net. We will choose the side lengths δ_1 and δ_2, later. Then using union bound over the grids in the net, probability in (26) can be bounded by

$$\sum_{l_1,l_2} \mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\mathcal{G}_{l_1,l_2}} \sum_{l=1}^{N_1(T_a^k)} \log\left(1-(X_l-m)\lambda_1-(\tilde{B}-f(X_l))\lambda_2\right) \geq \frac{N_1(T_a^k)g(T_a^k)}{(1+\eta)^{i+1}}, D_j, C_i \right). \tag{27}$$

On the set D_j, T_a^k is at least $\underline{t} = (1+\tilde{\eta})^{k-1}(1+\gamma)^j$. Thus, using monotonicity of $g(\cdot)$, the probability in the summation above can be bounded by

$$\mathbb{P}\left(\max_{\boldsymbol{\lambda}\in\mathcal{G}_{l_1 l_2}} \sum_{l=1}^{N_1(T_a^k)} \log\left(1-(X_l-m)\lambda_1-(\tilde{B}-f(X_l))\lambda_2\right) \geq \frac{N_1(T_a^k)g(\underline{t})}{(1+\eta)^{i+1}}, D_j, C_i \right). \tag{28}$$

Let the maximum in the expression above be attained at some point in the grid, say $(\lambda_1^*, \lambda_2^*)$. Furthermore, let $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ denote one of the corner points of the grid $\mathcal{G}_{l_1,l_2}$ such that $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ is in the interior of $\mathcal{R}(m, \tilde{B})$. Then, $1-(X_l-m)\lambda_1^*-(\tilde{B}-f(X_l))\lambda_2^*$ equals

$$1-(X_l-m)\tilde{\lambda}_1-(\tilde{B}-f(X_l))\tilde{\lambda}_2+(X_l-m)\left(\tilde{\lambda}_1-\lambda_1^*\right)+\left(\tilde{B}-f(X_l)\right)\left(\tilde{\lambda}_2-\lambda_2^*\right).$$

Since λ_1^* and $\tilde{\lambda}_1$ are in the same grid, they differ by at most δ_1. Similarly, $\tilde{\lambda}_2$ and λ_2^* differ by at most δ_2. Thus, the r.h.s. in the above expression can be upper bounded by

$$1-(X_l-m)\tilde{\lambda}_1-(\tilde{B}-f(X_l))\tilde{\lambda}_2+|X_l-m|\,\delta_1+\left|\tilde{B}-f(X_l)\right|\delta_2.$$

Let $Y_l := \log\left(1-(X_l-m)\tilde{\lambda}_1-(\tilde{B}-f(X_l))\tilde{\lambda}_2+|X_l-m|\,\delta_1+\left|\tilde{B}-f(X_l)\right|\delta_2\right)$. Clearly, Y_l are i.i.d. random variables. Let Y be independent and identically distributed as Y_l. The probability in (28) can then be bounded by

$$\mathbb{P}\left(\sum_{l=1}^{N_1(T_a^k)} Y_l \geq \frac{N_1(T_a^k)g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)^{i+1}}, D_j, C_i \right). \tag{29}$$

For $0 \leq \theta \leq 1$, let

$$\Lambda_Y(\theta) = \log \mathbb{E}\left(e^{\theta Y}\right) \quad \text{and} \quad \theta^* = \operatorname*{argmax}_{0\leq\theta\leq 1} \left\{ \frac{\theta g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)^{i+1}} - \Lambda_Y(\theta) \right\}.$$

Clearly, $\theta^* \geq 0$. Using Chernoff-like argument, we bound the expression in (29) by

$$\mathbb{P}\left(e^{\sum_{l=1}^{N_1(T_a^k)}(\theta^* Y_l - \Lambda_Y(\theta^*))} \geq e^{N_1(T_a^k)\left(\frac{\theta^* g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)^{i+1}} - \Lambda_Y(\theta^*)\right)}, D_j, C_i \right).$$

Observe that by the choice of θ^*, the term in the exponent in the r.h.s. above is positive. Thus the above probability is bounded by choosing lower bound for $N_1(T_a^k)$ by

$$\mathbb{P}\left(e^{\sum_{l=1}^{N_1(T_a^k)} (\theta^* Y_l - \Lambda_Y(\theta^*))} \geq e^{(1+\eta)^i \left(\frac{\theta^* g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)^{i+1}} - \Lambda_Y(\theta^*) \right)}, D_j, C_i \right),$$

which can further be bounded by

$$\mathbb{P}\left(e^{\mathbb{1}(C_i \cap D_j) \sum_{l=1}^{N_1(T_a^k)} (\theta^* Y_l - \Lambda_Y(\theta^*))} \geq e^{(1+\eta)^i \left(\frac{\theta^* g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)^{i+1}} - \Lambda_Y(\theta^*) \right)} \right).$$

From Markov's Inequality, and by the choice of θ^*, the above probability is less than

$$\mathbb{E}\left(e^{\mathbb{1}(C_i \cap D_j) \sum_{l=1}^{N_1(T_a^k)} (\theta^* Y_l - \Lambda_Y(\theta^*))} \right) e^{-(1+\eta)^i \max_{0 \leq \theta \leq 1} \left(\frac{\theta g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)^{i+1}} - \Lambda_Y(\theta) \right)},$$

which is less than

$$\mathbb{E}\left(e^{\sum_{l=1}^{N_1(T_a^k)} (\theta^* Y_l - \Lambda_Y(\theta^*))} \right) e^{-(1+\eta)^i \left(\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)^{i+1}} - \Lambda_Y(1) \right)}.$$

Notice that the term inside the expectation in the previous expression is a 1-mean martingale, and the other term is bounded by choosing $\theta = 1$. Thus, (29), and hence, (28) is bounded by

$$\exp\left(- \left(\frac{g\left((1+\tilde{\eta})^{k-1} (1+\gamma)^j \right)}{(1+\eta)} - (1+\eta)^i \Lambda_Y(1) \right) \right). \tag{30}$$

We now evaluate $\Lambda_Y(1)$ to simplify the above bound. Observe that

$$(1+\eta)^i \Lambda_Y(1) \leq (1+\eta)^i \log\left(1 - \epsilon_2 \tilde{\lambda}_1 - \epsilon_1 \tilde{\lambda}_2 + \mathbb{E}\left(|X_l - m|\right) \delta_1 + \mathbb{E}\left(\left| \tilde{B} - f(|X_l|) \right| \right) \delta_2 \right).$$

Using this in (30), probability in (28) is bounded by

$$e^{-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)}} \left(1 - \epsilon_2 \tilde{\lambda}_1 - \epsilon_1 \tilde{\lambda}_2 + \mathbb{E}\left(|X_l - m|\right) \delta_1 + \mathbb{E}\left(\left| \tilde{B} - f(|X_l|) \right| \right) \delta_2 \right)^{(1+\eta)^i}.$$

Using $1 + x \leq e^x$, the above expression is less than

$$e^{-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^j\right)}{(1+\eta)}} e^{-\epsilon_2 \tilde{\lambda}_1 (1+\eta)^i - \epsilon_1 \tilde{\lambda}_2 (1+\eta)^i + \mathbb{E}(|X_l - m|) \delta_1 (1+\eta)^i + \mathbb{E}\left(\left| \tilde{B} - f(|X_l|) \right| \right) \delta_2 (1+\eta)^i}. \tag{31}$$

Choosing δ_1 and δ_2 as follows:

$$\delta_1 = \frac{(1+\eta)^{-i}}{f^{-1}\left(\tilde{B}\right) - \tilde{m}}, \quad \& \quad \delta_2 = \frac{(1+\eta)^{-i}}{\tilde{B} - f\left(|\tilde{m}|\right)},$$

and substituting in (31), the probability in (28) can be bounded by

$$\exp\left(-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^{j}\right)}{(1+\eta)}\right)\exp\left(-\epsilon_2\tilde{\lambda}_1(1+\eta)^{i}-\epsilon_1\tilde{\lambda}_2(1+\eta)^{i}\right)C_2, \tag{32}$$

where,

$$C_2=\exp\left(\frac{\mathbb{E}_{\mu_1}\left(|X_l-m|\right)}{f^{-1}(\tilde{B})-m}+\frac{\mathbb{E}_{\mu_1}\left(\left|\tilde{B}-f(|X_l|)\right|\right)}{\tilde{B}-f\left(|m|\right)}\right).$$

Recall that $(\tilde{\lambda}_1,\tilde{\lambda}_2)$ is a corner point of the grid under consideration, and hence, $\tilde{\lambda}_1$ is either $l_1\delta_1$, or $(l_1+1)\delta_1$ and similarly, $\tilde{\lambda}_2\in\{l_2\delta_2,(1+l_2)\delta_2\}$. The above expression is further upper-bounded by choosing $\tilde{\lambda}_1=l_1\delta_1$ and $\tilde{\lambda}_2=l_2\delta_2$. Probability in (28) is bounded with these substitutions in (32) by

$$C_2\exp\left(-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^{j}\right)}{(1+\eta)}\right)\exp\left(-\frac{\epsilon_2 l_1}{f^{-1}\left(\tilde{B}\right)-m}-\frac{\epsilon_1 l_2}{\tilde{B}-f\left(|m|\right)}\right).$$

Using this bound in (27), (26) is bounded by

$$\sum_{l_1,l_2}C_2\exp\left(-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^{j}\right)}{(1+\eta)}\right)\exp\left(-\frac{\epsilon_2 l_1}{f^{-1}\left(\tilde{B}\right)-m}-\frac{\epsilon_1 l_2}{\tilde{B}-f\left(|m|\right)}\right).$$

Since the summation in the expression is finite for l_1 and l_2 ranging over all positive integers, on summing, we get the desired bound. ■

F.2. Proof of Proposition 20

Recall that for $\epsilon_2>0$, and $\gamma>0$, $m=m(\mu_1)-\epsilon_2$, and $J_M=\frac{\log\left(T/(1+\tilde{\eta})^{k-1}\right)}{\log(1+\gamma)}$ and for $j\in\{0,1,\cdots,J_M\}$, the event D_j is defined as

$$D_j=\left\{(1+\tilde{\eta})^{k-1}(1+\gamma)^{j}\le T_a^k\le(1+\tilde{\eta})^{k-1}(1+\gamma)^{j+1}\right\}.$$

Recall that for $t>0$, $\eta>0$, and $i\ge 0$, we had $C_i(t)=\left\{(1+\eta)^{i}\le N_1(t)\le(1+\eta)^{i+1}\right\}$.

Let $\eta_T=\delta_T'$, where recall that $\delta_t':=\log\left(1+\frac{1}{\log\log t}\right)$, and define $I_{M,j}=\frac{\log\left((1+\tilde{\eta})^{k-1}(1+\gamma)^{j+1}\right)}{\log(1+\eta_T)}$. Using union bound and Lemma 22 the required probability is bounded by

$$\sum_{j=0}^{J_M}\sum_{i=0}^{I_{M,j}}C_1\exp\left(-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^{j}\right)}{(1+\eta_T)}\right),$$

which equals

$$\sum_{j=0}^{J_M}I_{M,j}C_1\exp\left(-\frac{g\left((1+\tilde{\eta})^{k-1}(1+\gamma)^{j}\right)}{(1+\eta_T)}\right), \tag{33}$$

where C_1 is as defined in Lemma 22. Recall that $g(t) = \left(1 + \log\left(1 + \frac{1}{\log\log t}\right)\right)^2 \log(t)$. For $t = (1+\tilde{\eta})^{k-1}(1+\gamma)^j$, $t \leq T$ and $\delta'_t \geq \delta'_T$. Substituting for the function $g(\cdot)$, bounding the terms involving δ'_t by δ'_T, and using that for all $j \in [J_M]$ $\eta_{j,k} \geq \eta_{J_M,k} = \log\left(1 + \frac{1}{\log\log T}\right)$, the last expression in (33) is bounded by

$$\frac{C_1 (1+\tilde{\eta})^{-(k-1)\left(1+\delta'_T\right)}}{\log\left(1+\delta'_T\right)} \sum_{j=0}^{J_M} \frac{\log\left((1+\tilde{\eta})^{k-1}(1+\gamma)^{j+1}\right)}{(1+\gamma)^{j(1+\delta'_T)}}.$$

Bounding the above expression by summing for j ranging over all positive integers, we get the desired bound.