Regret Tail Characterization of Optimal Bandit Algorithms with Generic Rewards

Subhodip Panda, Shubhada Agrawal

IEEE International Symposium on Information Theory (ISIT 2026)

May 19, 2026

Table of Contents

Introduction

- Motivation
  - While expected regret is a canonical performance metric in stochastic bandits, it provides only a coarse summary of algorithmic behavior.
  - A more informative characterization is obtained by studying the distribution of the regret, and in particular its tail behavior.
  - While minimizing expected regret is the classical objective, recent work shows that even such algorithms can exhibit heavy regret tails, incurring large regret with non-negligible probability.
- Problem Statement
  - Previous work establishes regret tail bounds of optimal bandit algorithms for the single-parameter exponential family (SPEF) of reward distributions.
  - Can we characterize the regret-tail behavior of asymptotically optimal bandit algorithms for broader nonparametric classes of reward distributions?

Contributions

- We extend the $\mathrm{KL}_{\mathrm{inf}}$-UCB algorithm of *Agrawal et al.* to a much broader family of reward distributions $\mathcal{L}$ (to be introduced later) and show that it is optimal.
- Our main technical result is the regret tail upper bound for the generalized $\mathrm{KL}_{\mathrm{inf}}$-UCB algorithm. This is the first regret-tail upper bound for asymptotically optimal bandit algorithms beyond parametric models. The above result is tight under some additional structural assumptions.
- As immediate corollaries, we get the upper bounds for the classical settings: (a) the moment-bounded heavy-tailed class considered in *Agrawal et al. (2021a)* and (b) the bounded-support reward distributions.
- For the special yet nontrivial case of finitely-supported distributions, we provide a much tighter regret tail upper bound analysis, and show that it matches the lower bound of previous work.

Problem Setup

- We consider a bandit problem with K-arms indexed by $a \in [K] := \{1, \ldots, K\}$, with $K \geq 2$.
- $\mathcal{P}(\Re)$ denote the collection of all probability measures on $\Re$.
- $\mathcal{M} \subseteq \mathcal{P}(\Re)$, we denote a collection of bandit instances by $\mathcal{M}^K$, a collection of vectors of K distributions, each from $\mathcal{M}$.
- A bandit environment $\boldsymbol{\nu} \in \mathcal{M}^K$ such that $\boldsymbol{\nu} = \{\nu_1, \nu_2, \ldots, \nu_K\}$
- Let $\mu^* = \max\{\mu_a : a \in [K]\}$ be the expected reward of the optimal arm. The sub-optimality gap of an arm a as $\Delta_a = \mu^* - \mu_a$.
- We assume that arm 1 is the optimal arm, and means are ordered so that $\mu_1 > \mu_2 > \ldots > \mu_K$.

Preliminaries

- At each time t, the algorithm selects an arm $A_t \in [K]$ based on the past information, and receives a reward Y_t.
- The number of times each arm a is pulled till time t is denoted by $N_a(t) \triangleq \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$.
- In addition, for each arm a and all rounds t such that $N_a(t) \geq 1$, the empirical reward distribution of arm a is defined as $\hat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{I}\{A_s = a\}$.
- The quality of an algorithm π is evaluated using the standard notion of *expected regret*, defined next. The expected regret till time $T \geq 1$ is defined as

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

Preliminaries

- Lai-Robbins Lower Bound defines an optimal algorithm.

Definition

An algorithm is said to be $\mathcal{M}^K$-optimal if for every $\boldsymbol{\nu} \in \mathcal{M}^K$ and for each sub-optimal arm a, the following holds:

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

- More generally, for a probability distribution ν and $x \in \Re$,

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

- Here, $\mathrm{KL}(\nu_a, \nu') = \int \log \frac{d\nu_a}{d\nu'} d\nu_a$ is the Kullback-Leibler divergence between ν_a and ν'.

Distribution Class $\mathcal{L}$

1. *Agrawal et al., Lemma 1* show that it is necessary to impose certain restrictions on class $\mathcal{M}$, otherwise $\mathrm{KL}_{\inf}^{\mathcal{M}}(\cdot,\cdot) = 0$, which makes it impossible to achieve logarithmic regret.
2. Thus, we focus on a specific subclass $\mathcal{L}$ satisfying certain regularity properties as follows

Assumption

For $\nu \in \mathcal{L}$, let $\hat{\nu}_n$ be the empirical distribution of n i.i.d. samples from ν, having mean $m(\nu)$. Let $g(n) = C_1 \log(1+n) + C_2$ for some $C_1, C_2 > 0$. Then, $\nu, \mathcal{L}$ satisfy:

$$\mathbb{P}\left(\exists n \in \mathbb{N} : n\mathrm{KL}_{\inf}^{\mathcal{L}}(\hat{\nu}_n, m(\nu)) - g(n) \geq x\right) \leq e^{-x}. \tag{2}$$

Distribution Class $\mathcal{L}$

Assumption

For $\nu \in \mathcal{L}$, let $\hat{\nu}_n$ be the empirical distribution of n i.i.d. samples from ν, having mean $m(\nu)$, and let $\delta > 0$. Then there exist constants $d_0 > 0$ and $c_\nu > 0$ s.t. for all $d < d_0$ the following holds:

$$\mathbb{P}\left(\mathrm{KL}_{\text{inf}}^{\mathcal{L}}(\hat{\nu}_n, m(\nu) + \delta) \leq \mathrm{KL}_{\text{inf}}^{\mathcal{L}}(\nu, m(\nu) + \delta) - d\right) \leq e^{-nc_\nu d^2}. \tag{3}$$

- The above conditions put structural restrictions on the underlying distribution class $\mathcal{L}$.
- Several interesting model classes, such as moment-bounded and finite support distribution classes, follow the above assumptions.

Optimality of $\mathrm{KL}_{\mathrm{inf}}$-UCB Algorithm

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

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

Theorem

For $\boldsymbol{\nu} \in \mathcal{L}^K$ and for every suboptimal arm a, let $f_a(t) = \log(t) + 2\log\log(t) + g(N_a(t))$. Then Generalized $\mathrm{KL}_{\mathrm{inf}}$-UCB, with inputs $(K, f_a(.))$ satisfies

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

Regret Tail Lower Bound

Theorem

Let π be $\mathcal{M}^K$-optimal algorithm, and for $\gamma \in (0,1)$, let $\mathcal{D}_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$. Then, for $\boldsymbol{\nu} \in \mathcal{M}^K$, and for an i-th best arm,

$$\liminf_{T\to\infty} \inf_{x\in\mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}_\nu^\pi(N_i(T) > x)}{\log x} \geq -\sum_{j=1}^{i-1} \inf_{\substack{\widetilde{\nu}\in\mathcal{M}: \\ m(\widetilde{\nu})<\mu_i}} \frac{\mathrm{KL}(\widetilde{\nu}, \nu_j)}{\mathrm{KL}_{\mathsf{inf}}(\widetilde{\nu}, \mu_i)}. \tag{4}$$

- Large-regret events can still occur with a non-negligible probability.
- *Fan et al.* proved the above result for the special case when $\mathcal{M}^K$ is SPEF. However, we extend this lower bound result to any general class of distributions.

Regret Tail Upper Bounds

- Below theorem provides the first regret tail upper bound for asymptotically optimal bandit algorithms beyond SPEF models.

Theorem

Let π be $\mathcal{L}^K$-optimal Generalized $\mathrm{KL}_{\mathsf{inf}}$-UCB algorithm. For $\gamma \in (0,1)$, let $\mathcal{D}_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$. Then, for any $\boldsymbol{\nu} \in \mathcal{L}^K$, and for the i-th best arm in $\boldsymbol{\nu}$,

$$\limsup_{T\to\infty} \inf_{x\in\mathcal{D}_\gamma(T)} \frac{\log \mathbb{P}^\pi_{\boldsymbol{\nu}}(N_i(T) > x)}{\log x} \leq -(i-1). \tag{5}$$

- As corollaries, the above result extends to broad distribution classes such as moment-bounded and bounded supported distribution classes.
- The above upper bound is tight under the additional structural assumption of discrimination equivalence.

Regret Tail Upper Bounds

- Finitely supported class is not discrimination equivalent.
- However, we provide a much tighter analysis of this specific case given by the following theorem.

Theorem (Tight regret tail upper bound for finitely-supported class)

Let π be $\mathcal{L}^K_{[a,b],\mathcal{X}}$-optimal $\mathrm{KL}_{\mathrm{inf}}$*-UCB algorithm with* $f_a(t) = \log t + \log\log t$*. For $\gamma \in (0,1)$, let* $\mathcal{D}_\gamma(T) = [\log^{1+\gamma}(T), (1-\gamma)T]$*. Then, for each $\boldsymbol{\nu} \in \mathcal{L}^K_{[a,b],\mathcal{X}}$, and for the i-th best arm in $\boldsymbol{\nu}$,*

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

Conclusion and Future works

- We investigated the regret-tail behavior of stochastic multi-armed bandit algorithms that are optimal over a broad class of reward distributions, beyond SPEF classes.
- We extend the $\mathrm{KL}_{\mathsf{inf}}$-UCB algorithm to a substantially more general class of reward distributions and establish its asymptotic optimality over this class.
- We further demonstrate the tightness of the proposed regret-tail upper bounds.
- For the general distribution class $\mathcal{L}$, a gap persists between the regret-tail upper and lower bounds when additional structure is not present.
- Further, the study of fragility issues in the Thompson sampling remains a promising direction.