Regret-Minimization in Risk-Averse Bandits

Shubhada Agrawal
TIFR, Mumbai
shubhadaiitd@gmail.com

Sandeep Juneja
TIFR, Mumbai
juneja@tifr.res.in

Wouter M. Koolen
CWI, Amsterdam
wmkoolen@cwi.nl

***Abstract*—Classical regret minimization in a bandit framework involves a number of probability distributions or arms that are not known to the learner but that can be sampled from or pulled. The learner's aim is to sequentially pull these arms so as to maximize the number of times the best arm is pulled, or equivalently, minimize the regret associated with the sub-optimal pulls. Best is classically defined as the arm with the largest mean. Lower bounds on expected regret are well known, and lately, in great generality, efficient algorithms that match the lower bounds have been developed. In this paper we extend this methodology to a more general risk-reward set-up where the best arm corresponds to the one with the lowest average loss (negative of reward), with a multiple of Conditional-Value-at-Risk (CVaR) of the loss distribution added to it. CVaR is a popular tail risk measure. The settings where risk becomes an important consideration, typically involve heavy-tailed distributions. Unlike in most of the previous literature, we allow for all the distributions with a known uniform bound on the moment of order $(1+\epsilon)$, allowing for heavy-tailed bandits. We extend the lower bound of the classical regret minimization setup to this setting and develop an index-based algorithm. Like the popular KL-UCB algorithm for the mean setting, our index is derived from the proposed lower bound, and is based on the empirical likelihood principle. We also propose anytime-valid confidence intervals for the mean-CVaR trade-off metric. En route, we develop concentration inequalities, which may be of independent interest.**

I. Introduction

A multi-armed bandit (MAB) problem is a sequential decision-making problem in which a learner is presented with K unknown probability distributions, or *arms*, denoted by $\mu = (\mu_1, \ldots, \mu_K)$. At each time t, it samples or *pulls* one of the arms, A_t, and observes an independent sample, X_t, generated from the corresponding distribution, μ_{A_t}. These samples correspond to rewards, and the learner's aim is to maximize the cumulative average reward. This can also be modelled using a loss-version, i.e., when each sample is viewed as a loss. Here, the aim is to minimize the cumulative average loss. These objectives can be equivalently formulated as minimizing the cumulative average regret defined as the absolute difference between the average reward/loss gathered by learner, and the policy playing the best-arm (one with maximum mean-reward or minimum mean-loss) in all the rounds.

Typically, in both these formulations, metric associated with measuring the "goodness" or "badness" of an arm is mean of the associated distribution. However, in many applications, for example, in finance, health-care, insurance, etc., the expected reward/loss is typically not the primary desirable objective, as the learners are sensitive to the worst-case outcomes. In these settings, along with this average metric, it is important to understand the behavior of tails of the distributions (possibility of extreme events). In clinical trials, for example, there may be a high variability associated with a treatment, meaning that the treatment with good average performance may result in adverse outcomes for some patients. While investing in financial securities, an investor typically looks for a low-risk investment that also generates good returns. Thus, it is important to include a reasonable formulation of "tail risk" in the performance metric, together with the mean.

Tail risk has been an important topic in finance, and related fields. Popular measures of tail-risk include value-at-risk (VaR) and conditional value-at-risk (CVaR), which are widely used (see, [1], [2] for applications in finance and optimization). CVaR is a coherent risk measure that can be used to mathematically formalize the idea of risk-awareness, whence, preferred over VaR (see, [3] for definition of coherent risk measures). We therefore use CVaR as a measure of tail-risk in our formulation (see Section II for the definition and alternative formulations of CVaR).

The regret-minimization MAB problem with mean as the performance metric is well studied in literature. Tight lower bound on the expected regret, and algorithms asymptotically matching the lower bound, exist in literature. A lower bound on the problem was first proposed in [4] for parametric family of arm-distributions and later generalized in [5]. Since then, this problem with mean performance metric has been solved under different classes of allowed arm-distributions. See [6] for the popular UCB-1 algorithm for bounded support distributions, [7] for an optimal algorithm for single parameter family of distributions, [8] for bounded support distributions, and [9] for that in the heavy-tailed setting. Also see [10], [11] for a survey of the extensive literature on this problem and its variants.

Lately, there has been some interest in the MAB problems with the metric measuring the quality of an arm being a measure of tail-risk. For the best-arm identification (BAI) variant of the prolem, where the aim is to identify the arm with the best performance metric, see [12]–[14]. [15] recently proposed an optimal algorithm for this problem. [16], [17] study the regret-minimization variant with the CVaR metric, when the underlying distributions have a bounded support.

Typically, in applications, one is interested in maximizing the average return while minimizing the extreme losses.

Viewing the samples as losses, this translates to minimizing a conic combination of mean and CVaR. [18] and [15] studied the best-arm variant with this objective in great generality, allowing for arm-distributions to be heavy-tailed (i.e., having infinite moment generating function for all $\theta > 0$).

Applications where tail-risk becomes an important consideration typically involve distributions that are heavy tailed. Building upon recent works in [15] and [9], in this work, we consider the regret-minimization MAB problem with a mild restriction on the arm-distributions (allowing for heavy tails). The performance metric we consider is a conic combination of mean and CVaR, and is similar to that in [15]. We tackle the regret-minimization MAB problem with this risk-averse performance metric, i.e., the best arm is the one with the minimum value of a conic combination of mean and CVaR. We propose a lower bound on the average cumulative regret incurred by an algorithm (Theorem 1) and also develop an algorithm that we conjecture to be asymptotically optimal. See Assumption (A1) in Section IV, that has been shown to be true in the mean performance metric setting. [19] study a related problem, in which they include the risk-sensitivity, measured using CVaR, in the constraints instead of in the objective, thus formulating it as a constrained regret-minimization problem. We develop anytime-valid confidence intervals for the mean-CVaR performance metric that are based on the empirical likelihood method. Analysis of our algorithm and the proposed confidence intervals relies on new concentration inequalities (Proposition 3), that may be of independent interest.

Roadmap: The rest of the paper is structured as follows. In Section II, we first define VaR and CVaR and provide their alternative formulations that will be useful in our analysis and in the proposed algorithm. In this section we also formally introduce the problem setup and give the proposed lower bound on the quantity of interest. In Section III, we present our proposed index-based algorithm. Its theoretical guarantees are presented in Section IV. In this section, we also review the simpler dual representations for the KL-projection functions that appear in the lower bound. These are crucial for our algorithm and its analysis. Section V has our proposed anytime-valid confidence intervals for a conic combination of mean and CVaR. We also present some useful concentration inequalities in this section, and give a brief outline of their proofs. We conclude in Section VI.

II. Problem setup and Lower bound

In this section, we formally introduce the problem and give the necessary background. Let $\mathcal{P}(\Re)$ denote the collection of all probability measures on $\Re$. For $\eta \in \mathcal{P}(\Re)$, let $m(\eta)$ denote its mean. We denote the CDF function associated with η by $F_\eta(.)$, i.e., $F_\eta(y) = \eta((-\infty, y])$, for $y \in \Re$. Throughout this paper, samples will correspond to random losses from a loss-distribution (for example, from an investment).

VaR, CVaR: With this notation, for $\pi \in (0,1)$, VaR at level π, denoted as $x_\pi(\eta)$, represents the π^{th} quantile of η and is defined as

$$x_\pi(\eta) = \min\{z : F_\eta(z) \geq \pi\}.$$

CVaR at level π, denoted as $c_\pi(\eta)$, represents the average loss conditioned on losses being greater than $x_\pi(\eta)$. Formally,

$$c_\pi(\eta) := \frac{F_\eta(x_\pi(\eta)) - \pi}{1-\pi} x_\pi(\eta) + \frac{1}{1-\pi} \int_{x_\pi(\eta)}^{\infty} y dF_\eta(y).$$

It has many equivalent formulations (see [1] for a comprehensive tutorial), most relevant to us being the minimization form, given in (1) below. The minimizer in (1) is $x_\pi(\eta)$, the VaR at π.

$$c_\pi(\eta) = \min_{z \in \Re} \left\{ z + \frac{1}{1-\pi} \mathbb{E}_\eta \left((X - z)_+ \right) \right\}. \tag{1}$$

Observe from the above formulation that CVaR is min of linear functions of η. Whence, $c_\pi(\eta)$ is concave in η.

Recall that the arms correspond to loss-distributions. Hence the arm with a small mean as well as a small CVaR is preferable. Given a bandit instance, $\mu = (\mu_1, \ldots, \mu_K)$, for fixed constants $\alpha_1 > 0$, $\alpha_2 > 0$, and $\pi \in (0,1)$, we associate

$$o_\pi(\mu_a) := \alpha_1 m(\mu_a) + \alpha_2 c_\pi(\mu_a)$$

as a measure of "badness" of arm a.

Without loss of generality, for simplicity of presentation, we assume henceforth that arm 1 is the unique best arm, i.e.,

$$\mu_1 = \operatorname*{argmin}_{a \in [K]} o_\pi(\mu_a).$$

The average regret associated with pulling a sub-optimal arm $a \neq 1$ is denoted by

$$\Delta_a := o_\pi(\mu_a) - o_\pi(\mu_1).$$

Let $N_a(t)$ denote the number of pulls of arm a in time $t-1$. Then, the cumulative expected regret of the algorithm till time T, $\mathbb{E}(R_T)$, is given by

$$\mathbb{E}(R_T) := \sum_{a \neq 1} \mathbb{E}(N_a(t)) \Delta_a. \tag{2}$$

Notice that in order to bound $\mathbb{E}(R_T)$, it is sufficient to bound the average number of pulls for each of the sub-optimal arms.

As in the mean-regret-minimization setting (see [9]), it can be shown that without any restriction on the arm-distributions, the lower bound on expected regret incurred by an algorithm will be unbounded. We hence impose a very mild restriction on the allowed arm-distributions. In particular, we focus on the class $\mathcal{L}$ of all distributions with a known uniform bound on $(1+\epsilon)^{th}$ moment, for some $\epsilon > 0$.

For $\eta \in \mathcal{P}(\Re)$, let

$$\mathbb{E}_\eta |X|^{1+\epsilon} = \int_{\Re} |y|^{1+\epsilon} \, d\eta(y)$$

denote its $(1+\epsilon)^{th}$ moment. In particular,

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

for positive constants $\epsilon > 0$, and B. Observe that higher moments may not exist for distributions in this class and that it includes many heavy-tailed distributions.

This bound on $(1+\epsilon)^{th}$ moment restricts the possible values for $x_\pi(\cdot)$, $c_\pi(\cdot)$, and $o_\pi(\cdot)$. In particular, define

$$Z := \left[-\left(B\pi^{\text{-}1}\right)^{\frac{1}{1+\epsilon}}, \left(B\left(1-\pi\right)^{\text{-}1}\right)^{\frac{1}{1+\epsilon}} \right],$$

$$C := \left[-B^{\frac{1}{1+\epsilon}}, \left(B\left(1-\pi\right)^{\text{-}1}\right)^{\frac{1}{1+\epsilon}} \right],$$

$$O := \left[-B^{\frac{1}{1+\epsilon}}(\alpha_1+\alpha_2), B^{\frac{1}{1+\epsilon}}c \right],$$

for some $c > 0$. Then, it can be shown that for distributions $\eta \in \mathcal{L}$, $x_\pi(\eta) \in Z$, $c_\pi(\eta) \in C$, and $o_\pi(\eta) \in O$. See [15] for a proof.

Next, for $\eta \in \mathcal{P}(\Re)$ and $x \in \Re$, we define the KL-projection function,

$$\mathrm{KL}_{\inf}(\eta, x) := \inf \left\{ \mathrm{KL}(\eta, \kappa) : \ \kappa \in \mathcal{L}, \ o_\pi(\kappa) \leq x \right\}. \quad (3)$$

See [5], [8], [20] for related quantities in the mean setting. This, and the related KL-projection functionals introduced later, will be crucial for the proposed lower bound, our algorithm and its analysis, and the proposed confidence intervals. Let us look at some of its properties.

Observe that for $\eta \in \mathcal{L}$ and $x \geq o_\pi(\eta)$, $\mathrm{KL}_{\inf}(\eta, x)$ equals 0 as η is a feasible solution to the $\mathrm{KL}_{\inf}$ optimization problem. Moreover, it is monotonically non-increasing in the second argument for $x \leq o_\pi(\eta)$.

Recall that $c_\pi(\eta)$ is concave in η. This follows from the formulation in (1) which shows that $c_\pi(\cdot)$ is a minimum of linear functions. Thus, the mean-CVaR constraint, i.e., $o_\pi(\kappa) \leq$ constraint, in $\mathrm{KL}_{\inf}$ renders the corresponding optimization problem non-convex.

One can similarly define $\widetilde{\mathrm{KL}_{\inf}}$, with $o_\pi(\kappa) \geq x$ constraint, i.e.,

$$\widetilde{\mathrm{KL}_{\inf}}(\eta, x) := \inf \left\{ \mathrm{KL}(\eta, \kappa) : \ \kappa \in \mathcal{L}, \ o_\pi(\kappa) \geq x \right\}. \quad (4)$$

However, this optimization problem is a convex optimization problem. Furthermore, for $\eta \in \mathcal{L}$ and $x \leq o_\pi(\eta)$, $\widetilde{\mathrm{KL}_{\inf}}(\eta, x) = 0$, and it is monotonically non-decreasing in the second argument for $x \geq o_\pi(\eta)$.

One can similarly define $\mathrm{KL}_{\inf}$ functionals with either $m(\eta)$ or $c_\pi(\eta)$ less than or greater than constraints. While these functionals with the mean constraints are symmetric (see, [20]), they are not symmetric for the CVaR case. Whence, $\mathrm{KL}_{\inf}$ and $\widetilde{\mathrm{KL}_{\inf}}$ are not symmetric. In particular, as observed earlier, while $\widetilde{\mathrm{KL}_{\inf}}$ is a convex optimization problem, $\mathrm{KL}_{\inf}$ is not. See [15] for a similar discussion in the CVaR setting. Both $\mathrm{KL}_{\inf}$ and $\widetilde{\mathrm{KL}_{\inf}}$ will be crucial to our analysis.

Extending the lower bound for the mean-regret-minimization problem to the mean-risk setting, using the change of measure argument, we get the following lower bound for our problem.

Theorem 1 (Lower bound): Fix $\pi \in (0,1), \alpha_1 > 0, \alpha_2 > 0$. Let $\mathcal{A}$ be an algorithm which, when acted on $\mu \in \mathcal{L}^K$, satisfies

$$R_T = o(T^\gamma), \quad \forall \gamma > 0.$$

Then,

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left(N_a(T)\right)}{\log T} \geq \frac{1}{\mathrm{KL}_{\inf}(\mu_a, o_\pi(\mu_1))}.$$

The proof of Theorem 1 above, uses standard change of measure arguments. It can also be proven using the Transportation Lemma of [21]. This proof is similar to those developed in [4], [5], and is omitted for space considerations.

Using (2) and Theorem 1, we get a lower bound on $\mathbb{E}\left(R_T\right)$.

Next, building upon the asymptotically optimal algorithm of [9] in the mean-setting, we porpose an algorithm which we conjecture to be asymptotically optimal. We present the proposed algorithm and its analysis first. We also develop concentration inequalities for empirical versions of $\mathrm{KL}_{\inf}$ and $\widetilde{\mathrm{KL}_{\inf}}$. These are used in developing anytime-valid confidence intervals for the conic combination of mean-CVaR under consideration and may be of independent interest. These are presented towards the end.

III. The algorithm: RA-$\mathrm{KL}_{\inf}$-LCB

The algorithm we propose is a UCB-like algorithm that derives its index using the $\mathrm{KL}_{\inf}$-functional that appears in the lower bound. KL-based UCB algorithms have been shown to be optimal in the mean-regret setting (see, [7], [9]). Our algorithm differs from those in literature in 2 ways. First, we work with a practically more relavant risk-averse objective. Second, in our setting, the samples correspond to losses. Hence we construct a lower confidence bound (LCB) on the true mean-risk measure, $o_\pi(\mu_a)$, for each arm a, and at each time, select the arm with the minimum value of the corresponding LCB.

For $\mu \in \mathcal{L}^K$, let $\hat{\mu}_a(n)$ denote the empirical distribution corresponding to $N_a(n)$ samples from arm a, and let $L_a(n)$ denote the LCB associated with arm a at time n. For thresholds $g_a(n)$:

$$g_a(n) = \log n + 2\log\log n + 2\log(1 + N_a(n)) + 1,$$

define

$$L_a(n) := \inf \left\{ x \in \Re : N_a(n)\, \mathrm{KL}_{\inf}(\hat{\mu}_a(n), x) \leq g_a(n) \right\}.$$

Using the definition of $\mathrm{KL}_{\inf}$, $L_a(n)$ can be reformulated as

$$\inf \left\{ o_\pi(\kappa) : \kappa \in \mathcal{L}, \ N_a(n)\, \mathrm{KL}(\hat{\mu}_a(n), \kappa) \leq g_a(n) \right\}.$$

In particular, our lower confidence bound for $o_\pi(\mu_a)$ at time n is the minimum mean-risk metric of distributions in $\mathcal{L}$ which are within a KL-ball of the empirical distribution, $\hat{\mu}_a(n)$.

The proposed algorithm, 'RA-$\mathrm{KL}_{\inf}$-LCB' takes as input the number of arms (K), the class-parameters (B, ϵ), and

the threshold functions $(g_a(\cdot))$. It begins by pulling each arm once. At time n, it computes $L_a(n)$ for each arm, and pulls

$$A_n \in \underset{a\in[K]}{\operatorname{argmin}}\ L_a(n),$$

breaking ties randomly. It then sets

$$N_{A_n}(n) \leftarrow N_{A_n}(n-1)+1,$$

and updates $\hat{\mu}_a(n)$, for each arm.

Computing the index: [15, Appendix H.1] give an alternate min-max formulation for $\mathrm{KL}_{\inf}$, which is related to its Lagrangian dual. We review these in Section IV. The inner maximization problem in this represenstation is a convex-optimization problem over 2 variables, and can be solved using, for example, ellipsoid method. The outer minimization problem, which corresponds to that in (1), may not be convex. Further work is needed to understand this optimization problem.

IV. Theoretical guarantees

Assumption (A1): For $a \neq 1$, the following summation is $o(\log T)$:

$$\sum_{n=1}^{T} \mathbb{P}\left(\mathrm{KL}_{\inf}(\hat{\mu}_a(n), o_\pi(\mu_1)) \leq \mathrm{KL}_{\inf}(\mu_a, o_\pi(\mu_1)) - \delta\right).$$

It has been shown to be true in the mean setting (see [9, Lemma 6] and [8, Theorem 12]). We expect this to be true although we do not have a proof at this time.

Conjecture 2 (Asymptotic optimality): Let $T > K \geq 2$, $\mu \in \mathcal{L}^K$, and

$$g_a(n) = \log n + 2\log\log n + 2\log(1+N_a(n)) + 1.$$

Then, under Assumption (A1), for each sub-optimal arm a,

$$\limsup_{T\to\infty} \frac{\mathbb{E}\left(N_a(T)\right)}{\log T} \leq \frac{1}{\mathrm{KL}_{\inf}(\mu_a, o_\pi(\mu_1))}.$$

Proof for Theorem 2 proceeds by analysing the events leading to selection of a sub-optimal arm. This can occur either due to index of the sub-optimal arm taking a lower deviation, or that for the optimal arm taking a higher deviation. These can be viewed as deviations of either $\mathrm{KL}_{\inf}(\hat{\mu}_a(n), o_\pi(\mu_1))$ or $\mathrm{KL}_{\inf}(\hat{\mu}_1(n), o_\pi(\mu_1))$.

The index for a sub-optimal arm can be small until it has been pulled sufficient number of times. This contributes to the main term in the regret. Assumption (A1) essentially says that the index for the sub-optimal arm evaluating to something below $o_\pi(\mu_1)$ is a rare event. Proposition 3 below shows that the RA-$\mathrm{KL}_{\inf}$-LCB index for optimal arm evaluating higher than the true value, $o_\pi(\mu_1)$ is rare, i.e., our index is a high probability lower bound on the true value of mean-CVaR metric of the associated distribution. In fact, we show later in Section V, that anytime valid confidence intervals for the conic combination of mean and CVaR can be constructed using the $\mathrm{KL}_{\inf}$ and $\widetilde{\mathrm{KL}_{\inf}}$ functionals.

Let us now review alternate representations for both these functionals. These correspond to their dual formulations, and will be used later in our analysis. We borrow these from [15, Appendix H], and state here for completeness.

For $\eta \in \mathcal{P}(\Re)$, $x \in O^o$,

$$\mathrm{KL}_{\inf}(\eta, x) \;=\; \min_{z\in Z} \max_{\boldsymbol{\lambda}\in\mathcal{R}(x,z)} \mathbb{E}_\eta\left(\log f(X, \boldsymbol{\lambda}, z, x)\right), \quad (5)$$

where $f(X, \boldsymbol{\lambda}, z, x)$ equals

$$1-(B-|X|^{1+\epsilon})\lambda_1-\lambda_2 x+\alpha_1\lambda_2 X+\alpha_2\lambda_2 z+\frac{\lambda_2\alpha_2}{1-\pi}(X-z)_+,$$

and $\mathcal{R}(x,z) \subset \Re^2$ is

$$\mathcal{R}(x,z) = \left\{\lambda_1 \geq 0, \lambda_2 \geq 0,\ \min_{y\in\Re} f(y, \boldsymbol{\lambda}, z, x) \geq 0\right\}.$$

Recall that $\mathrm{KL}_{\inf}$ optimization problem is non-convex. However, using formulation (1) for CVaR in the $o_\pi(\cdot)$ constraint, and writing the dual for the optimization problem with fixed z, we get the inner maximization problem in (3). The outer minimization corresponds to that in the minimization formulation for CVaR in (1). See [15] for details.

Similarly,

$$\widetilde{\mathrm{KL}_{\inf}}(\eta, x) \;=\; \max_{\boldsymbol{\rho}\in\boldsymbol{\mathcal{S}}(x)} \mathbb{E}_\eta\left(\log g(X, \boldsymbol{\rho})\right), \quad (6)$$

where $g(X, \boldsymbol{\rho})$ equals

$$1-\rho_1(B-|X|^{1+\epsilon})+\rho_2(x-\alpha_1 X) - \rho_4(1-\pi) - \left(\frac{\rho_2\alpha_2 X}{1-\pi}-\rho_4\right)_+,$$

and $\mathcal{S}(x) \subset \Re^3$ is

$$\mathcal{S}(x) = \left\{\rho_1 \geq 0, \rho_2 \geq 0, \rho_3 \in \Re :\ \min_{y\in\Re} g(y, \boldsymbol{\rho}) \geq 0\right\}.$$

These dual formulations for the two KL-projection functionals are crucial in proving the concentration inequalities proposed in the next section.

V. Confidence interval for $o_\pi(\cdot)$

We propose a KL-based anytime-valid confidence intervals for a conic combination of mean and CVaR for distributions in $\mathcal{L}$. Our confidence intervals are derived from the $\mathrm{KL}_{\inf}$ functional that appears in the lower bound, and a related $\widetilde{\mathrm{KL}_{\inf}}$ functional.

Let $\delta > 0$, $\eta \in \mathcal{L}$, and let $\hat{\eta}_n$ denote the empirical distribution corresponding to n samples from η. Let

$$L^n_{\hat{\eta}} = \inf\{x : n\,\mathrm{KL}_{\inf}(\hat{\eta}_n, x) \leq \log\delta^{-1} + 2\log(n+1) + 1\}$$

and

$$U^n_{\hat{\eta}} = \sup\{x : n\widetilde{\mathrm{KL}_{\inf}}(\hat{\eta}_n, x) \leq \log\delta^{-1} + 3\log(n+1) + 1\}.$$

Observe that the event $\left\{o_\pi(\eta) \leq L^n_{\hat{\eta}_n}\right\}$ is same as

$$\left\{n\,\mathrm{KL}_{\inf}(\hat{\eta}_n, o_\pi(\eta)) \geq \log\delta^{-1} + 2\log(n+1) + 1\right\},$$

and $\left\{o_\pi(\eta) \geq U^n_{\hat{\eta}_n}\right\}$ equals

$$\left\{n\widetilde{\mathrm{KL}_{\inf}}(\hat{\eta}_n, o_\pi(\eta)) \geq \log \delta^{-1} + 3\log(n+1) + 1\right\}.$$

Proposition 3: For $\eta \in \mathcal{L}$ and $\delta > 0$,

$$\mathbb{P}\left(\exists n : n\,\mathrm{KL}_{\inf}(\hat{\eta}_n, o_\pi(\eta)) \geq \log \delta^{-1} + 2\log(n+1) + 1\right)$$

is at most δ, and

$$\mathbb{P}\left(\exists n : n\widetilde{\mathrm{KL}_{\inf}}(\hat{\eta}_n, o_\pi(\eta)) \geq \log \delta^{-1} + 3\log(n+1) + 1\right)$$

is bounded by δ.

It follows from Proposition 3 that $o_\pi(\eta) \in [L^n_{\hat{\eta}_n}, U^n_{\hat{\eta}_n}]$ with probability at least $1-2\delta$. Moreover, $[L^n_{\hat{\eta}_n}, U^n_{\hat{\eta}_n}]$ are anytime-valid confidence intervals.

Proof of Proposition 3 relies on constructing mixtures of non-negative super-martingales. To see this, let $X_1, \ldots, X_n$ denote i.i.d. samples from η. Using the dual formulation for $\mathrm{KL}_{\inf}$ from (3), and using a sub-optimal choice ($= x_\pi(\eta)$) for the outer minimization, we have $n\,\mathrm{KL}_{\inf}(\hat{\eta}_n, o_\pi(\eta))$ is at most

$$\max_{\boldsymbol{\lambda} \in \mathcal{R}(o_\pi(\eta), x_\pi(\eta))} \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\lambda}, x_\pi(\eta), o_\pi(\eta)).$$

Since $\mathcal{R}(o_\pi(\eta), x_\pi(\eta))$ is compact (see [15, Appendix H]), and

$$g_i(\boldsymbol{\lambda}) := \log f(X_i, \boldsymbol{\lambda}, x_\pi(\eta), o_\pi(\eta))$$

satisfies the conditions of [15, Lemma E.1], we have that $n\,\mathrm{KL}_{\inf}(\hat{\eta}_n, o_\pi(\eta))$ is at most

$$\log \mathbb{E}_{\boldsymbol{\lambda} \sim q}\left(\prod_{i=1}^{n} f(X_i, \boldsymbol{\lambda}, x_\pi(\eta), o_\pi(\eta))\right) + 2\log(n+1) + 1.$$

Moreover, since $\eta \in \mathcal{L}$, it is easy to see that $\mathbb{E}_\eta\left(f(X_i, \boldsymbol{\lambda}, x_\pi(\eta), o_\pi(\eta))\right)$ is at most 1, and hence, exponential of the first term in the above bound is a mixture of non-negative super-martingales. Using this bound in the first expression in Proposition 3, exponentiating, and using Ville's inequality, we get the desired bound of δ. Bound for $\widetilde{\mathrm{KL}_{\inf}}$ can be established similarly.

VI. Conclusion

We consider regret-minimization problem with the metric of performance being a conic combination of mean and CVaR, a tail-risk measure. Since tail-risk becomes an important consideration in settings where the underlying distributions have heavy-tails, we allow for a very general class of arm-distributions. We propose a lower bound on cumulative expected regret of an algorithm, and develop an index-based algorithm that derives its index from the lower bound. This is similar to the optimal KL-UCB algorithm in a much simpler mean setting. We conjecture it to be optimal in the setting considered in this paper.

Typically, analysis of regret-minimization algorithms relies on controlling 2 types of deviations, one for the best-arm, and second for the sub-optimal arms. This requires developing new concentration inequalities. While our concentration inequalities in Proposition 3 handle one of these, we assume that the other deviation is rare. It has been established to be true in the mean-setting. Under this assumption, it can be shown that the proposed algorithm is asymptotically optimal. We also develop anytime-valid confidence intervals for the mean-CVaR metric using the functionals of probability measures that appear in the lower bound for the regret-minimization problem.

An important next step is proving the Assumption A1. It requires developing concentration inequality for min-max of empirical averages. Computing RA-$\mathrm{KL}_{\inf}$-LCB index may be time consuming. Hence, understanding the structure and solutions of $\mathrm{KL}_{\inf}$ optimization problem can be useful in developing both, the concentration inequalities for min-max, as well as efficient algorithms for computing the index.

Acknowledgment

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

References

[1] S. Sarykalin, G. Serraino, and S. Uryasev, "Value-at-risk vs. conditional value-at-risk in risk management and optimization," in *State-of-the-art decision-making tools in the information-intensive age*. Informs, 2008, pp. 270–294.

[2] R. T. Rockafellar, "Coherent approaches to risk in optimization under uncertainty," in *OR Tools and Applications: Glimpses of Future Technologies*. Informs, 2007, pp. 38–61.

[3] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical finance*, vol. 9, no. 3, pp. 203–228, 1999.

[4] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4 – 22, 1985.

[5] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for sequential allocation problems," *Advances in Applied Mathematics*, vol. 17, no. 2, pp. 122 – 142, 1996.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, May 2002.

[7] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz *et al.*, "Kullback–Leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, vol. 41, no. 3, pp. 1516–1541, 2013.

[8] J. Honda and A. Takemura, "Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3721–3756, 2015.

[9] S. Agrawal, S. Juneja, and W. M. Koolen, "Regret minimization in heavy-tailed bandits," *arXiv preprint arXiv:2102.03734*, 2021.

[10] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[11] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[12] J. Y. Yu and E. Nikolova, "Sample complexity of risk-averse bandit-arm selection," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[13] P. L.A., K. Jagannathan, and R. Kolla, "Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 5577–5586. [Online]. Available: http://proceedings.mlr.press/v119/l-a-20a.html

[14] Y. David and N. Shimkin, "Pure exploration for max-quantile bandits," in *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2016, pp. 556–571.

[15] S. Agrawal, W. M. Koolen, and S. Juneja, "Optimal best-arm identification methods for tail-risk measures," *arXiv preprint arXiv:2008.07606*, 2020.

[16] D. Baudry, R. Gautron, E. Kaufmann, and O.-A. Maillard, "Thompson sampling for cvar bandits," *arXiv preprint arXiv:2012.05754*, 2020.

[17] A. Tamkin, R. Keramati, C. Dann, and E. Brunskill, "Distributionally-aware exploration for cvar bandits," in *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

[18] A. Kagrecha, J. Nair, and K. Jagannathan, "Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 11 272–11 281.

[19] ——, "Constrained regret minimization for multi-criterion multi-armed bandits," *arXiv preprint arXiv:2006.09649*, 2020.

[20] S. Agrawal, S. Juneja, and P. Glynn, "Optimal δ-correct best-arm selection for heavy-tailed distributions," in *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 117. PMLR, 08 Feb–11 Feb 2020, pp. 61–110.

[21] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.