Optimal Best-Arm Identification Methods for Tail-Risk Measures

Anonymous Author(s)
Affiliation
Address
email

Abstract

Conditional value-at-risk (CVaR) and value-at-risk (VaR) are popular tail-risk measures in finance and insurance industries as well as in highly reliable, safety-critical uncertain environments where often the underlying probability distributions are heavy-tailed. We use the multi-armed bandit best-arm identification framework and consider the problem of identifying the arm from amongst finitely many that has the smallest CVaR, VaR, or weighted sum of CVaR and mean. The latter captures the risk-return trade-off common in finance. Our main contribution is an optimal δ-correct algorithm that acts on general arms, including heavy-tailed distributions, and matches the lower bound on the expected number of samples needed, asymptotically (as δ approaches 0). The algorithm requires solving a non-convex optimization problem in the space of probability measures, that requires delicate analysis. En-route, we develop new non-asymptotic empirical likelihood based concentration inequalities for tail-risk measures which are tighter than those for popular truncation-based empirical estimators.

1 Introduction

Tail risk is a common term used to quantify the losses occuring due to rare events, and has been an important topic in finance, insurance and other safety critical uncertain environments. [44] first formalized the problem of identifying optimal investment in financial assets as a multi-criteria optimization problem of maximizing the average return, while minimizing the risk (measured via variance). Since then, several other risk measures have been considered. Lately, risk-measures based on tails of the distribution, like the conditional value-at-risk (CVaR) and value-at-risk (VaR), have gained popularity in financial regulations and risk management (see, [48, 47]), where the underlying probability distributions are mostly heavy tailed (i.e. having infinite moment generating function for all $\theta > 0$). Also see [50, 46] for applications of CVaR and VaR in finance and optimization. As opposed to VaR, CVaR is a coherent risk-measure, and is a preferable metric (see, [4] for precise definition and properties of coherence). Outside finance, these tail-risk measures are being used to control risk in operations management, for example, in inventory management [3], in the supply chain management [51], etc. Recently, coherent risk measures, especially CVaR, have also been used in connection with fairness in machine learning [57].

The importance of these risk measures in the sequential decision making set-up has well been acknowledged (see, [49, 42]). Typically in the stochastic multi-armed bandit (MAB) literature, the "goodness" of an arm is measured using its mean. Tight asymptotic and finite time guarantees exist for different MAB problems with performance measured via mean objective (see, [27, 36, 15, 12, 1, 6, 53]). Also, see [11] for a survey of the variants of stochastic MAB problems. However, maximizing the average reward is not always the primary desirable objective. In clinical trials, for example, the treatment that is good on average might result in adverse outcomes for some patients. In finance, one is typically interested in balancing the mean return with the risk of extreme losses. Risk sensitivity

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

has been well studied in online learning setting, where in each round, the player sees reward from every arm (see, [24, 55]). However, there is very limited work which incorporates these risk-measures into the MAB framework.

In this paper, we provide a systematic approach for identifying the distribution (or arm) from a given finite set of distributions (or arms) with minimum tail-risk (as measured by CVaR or VaR, or by a conic combination of mean and CVaR, which we will henceforth refer to as the "mean-CVaR" objective). Adopting the best-arm identification (BAI) framework of the stochastic MAB problem, we consider algorithms that generate samples from the given arms, and are δ-correct, i.e., identify the correct answer (arm with minimum VaR, CVaR or mean-CVaR) with probability at least $1-\delta$, for some pre-specified confidence level δ. While ensuring δ-correctness, the aim is to minimize the number of samples needed by the algorithm before its termination. This is the typical fixed-confidence setting of the BAI MAB problem (see, [37, 2]). Variants of this problem have been widely studied in the literature, where the best-arm is the one with maximum mean (see, [43, 25, 5, 13, 26, 33, 27, 34]).

A relaxation of the pure exploration setting described above is the (ϵ, δ)-PAC setting, where the aim is to output an ϵ-optimal arm (for an appropriate notion of ϵ-optimality), with probability at least $1-\delta$, while minimizing the number of samples generated. [58, 17, 32] consider the pure exploration problem of identifying the arm with minimum risk in the (ϵ, δ)-PAC setting. While [58] consider both VaR and CVaR as measures of risk, [17, 32] focus on the VaR-problem. Recently, [39] and [35] have studied the BAI MAB problem with CVaR and mean-CVaR objectives, respectively, in the closely related "fixed-budget" framework, in which the total number of times the algorithm is allowed to generate samples is fixed, and the aim of the algorithm is to minimize the error-probability.

Contributions: We first solve our minimum tail risk identification problems in the simple commonplace setting of arm-distributions belonging to a canonical single parameter exponential family (SPEF) of distributions. Each distribution in this family is uniquely identified with its parameter. We show that both CVaR and VaR are monotonic functions of this parameter, as is the mean. Hence, finding the best-(CVaR/VaR/mean-CVaR) arm reduces to finding the arm with the minimum mean. Since risk-sensitive objectives are particularly important when there is a non-trivial probability of occurrence of extreme outcomes, it is important to consider arm-distributions beyond canonical SPEF, for which the above-mentioned equivalence breaks. We solve the VaR problem for arbitrary arm distributions. In contrast the CVaR problem is unlearnable in full generality: on the class of all arm distributions, any δ-correct algorithm requires an infinite number of samples in expectation to identify the best arm amongst any finite collection of arms (Remark 3.1). To avoid this, we impose a mild and standard raw $(1+\epsilon)$-moment restriction on the arm-distributions. Let $\mathcal{P}(\Re)$ denote the collection of all the probability distributions on the reals $\Re$, and let B and ϵ be positive constants. For risk measure CVaR and for the mean-CVaR objective, we restrict the class of allowed arm distributions to

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

Regarding the choice of ϵ and B: Firstly, BAI problems are important in simulation where the best model may need to be identified amongst many intricate models in terms of a performance measure such as CVaR or VaR, using minimal computational effort (see, [31]). Input distributions in simulation are known and may often involve heavy tails. E.g., by the use of Lyapunov-function-based techniques, bounds on moments of output random variables, B, can be determined. (see, e.g., [28] and references therein). Secondly, consider rewards (returns) from a number of hedge funds. Each time some amount of money is invested into a fund, a random return may be revealed from that fund but not from others. To assume that these returns come from a class of parametric distributions or have known bounded support is likely a substantially inaccurate simplification. Typically, from historical analysis, it is known that the distribution of securities have a particular tail index, say, $(1+\epsilon)$. For stock returns, extensive research suggests that $(1+\epsilon) \in [2, 5]$. For daily exchange rates and income and wealth distributions we may have $(1+\epsilon) \in (1, 2]$. Extreme value theory, under reasonable dependence structure amongst underlying securities, shows that a portfolio (a weighted sum) will also have the same tail index of $(1+\epsilon)$ (see, [18]). So the key approximation needed is in arriving at B. It is easy to arrive at distributions η and κ whose $(1+\epsilon)^{th}$ moments are arbitrarily far while the KL distance between them is arbitrarily close to zero. This makes it difficult to infer B from a given sample of data without further restriction on the two distributions. One may take a pragmatic view and approximate B by estimating the $(1+\epsilon)^{th}$ moment from observed samples and padding it up with a reasonably large factor. A further set of distributional assumptions would be needed to justify the above procedure to arrive at B. Again, verifying those assumptions will entail

similar problems. In practice, one may live with the above approximation even though in rare settings it may be inaccurate and lead to sub-optimal allocations in our algorithm. One accepts this risk as one often accepts the assumption that the distributions of the random samples from each arm are time stationary or are independent, even though these may only approximately be correct.

Key contribution: We simplify the lower bound characterizations on sample complexity of any δ-correct algorithm, and use them to develop a δ-correct algorithm whose sample complexity exactly matches the lower bound as $\delta \to 0$, for CVaR, mean-CVaR, and VaR problems. The mean-CVaR problem is conceptually and technically similar to the CVaR problem. Hence, for simplicity of presentation, we primarily focus on the CVaR setting in the main text and give details of the mean-CVaR setting in Appendix H. We also spell out the somewhat analogous analysis for the VaR setting towards the end (Section 4.2), with details deferred to Appendix G.

As is well known in the BAI MAB literature, the lower bound problem takes underlying arm distributions as inputs and solves for optimal weights that determine the proportion of samples that should be ideally allocated to each arm. The proposed algorithm uses a plugin strategy that at each sequential stage, modulo mild forced exploration, uses the generated empirical distributions as a proxy for the true distributions and arrives at weights that guide the sequential sampling strategy.

Given η_1, η_2 in $\mathcal{P}(\Re)$, let $\mathrm{KL}(\eta_1, \eta_2)$ denote the KL-divergence between them, i.e., $\mathrm{KL}(\eta_1, \eta_2) := \int \log \frac{d\eta_1}{d\eta_2}(y) d\eta_1(y)$. Furthermore, let $c_\pi(\eta)$ and $x_\pi(\eta)$ denote the CVaR and VaR at the given confidence level $\pi \in (0, 1)$, for the probability measure η (see Section 2 for definitions of these risk measures).

CVaR problem: Given $\eta \in \mathcal{P}(\Re)$ and $x \in \Re$, define functionals $\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}} : \mathcal{P}(\Re) \times \Re \longrightarrow \Re^+$, and $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}} : \mathcal{P}(\Re) \times \Re \longrightarrow \Re^+$, where $\Re^+$ denotes the non-negative reals, as

$$\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}}(\eta, x) := \min_{\kappa \in \mathcal{L}:\ c_\pi(\kappa) \geq x} \mathrm{KL}(\eta, \kappa) \quad \text{and} \quad \mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}(\eta, x) := \min_{\kappa \in \mathcal{L}:\ c_\pi(\kappa) \leq x} \mathrm{KL}(\eta, \kappa). \tag{1}$$

See, [2, 30, 14] for related quantities. These projection functionals appear in the lower bound (Section 3), and are central to our plugin algorithm.

Unlike in the mean case, $\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}}$ and $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}$ in (1) are not symmetric, and need to be studied separately. In particular, $\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}}$ is a convex optimization problem, while $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}$ is not. This is because $c_\pi(.)$ is a concave function, whence, the CVaR constraint in the $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}$ problem in (1) renders the feasible region non-convex (see Section 2). CVaR can be expressed as the optimal value of a minimization problem. This helped in re-expressing $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}$ as minimization over 2 variables, fixing one of which resulted in convex optimization over the other (see Section 3).

For proving δ-correctness, we develop a new concentration inequality for weighted sums of these functionals (Proposition 4.2). Dual representations of these suggest natural candidates for super-martingales, whose mixtures help us in proving the concentration result. Similar inequalities were developed in [38, 19, 54] in different settings. See [40, Chapter 20] for an overview of the method of mixtures. We also propose $\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}}$- and $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}$-based tight anytime-valid confidence intervals for CVaR for heavy-tailed distributions, and show that our intervals typically have smaller width compared to those for the popular truncation-based empirical estimators (see Section 4.3).

Since the distributions are no longer characterized by parameters, we work in the space of probability measures instead of in the Euclidean space. A key and non-trivial requirement for the proof of asymptotic optimality of the algorithm is the joint continuity of $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}$ and $\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}}$ in a well-chosen metric, which should generate a topology that is sufficiently fine to ensure this continuity, but coarse to ensure fast convergence of the empirical distributions to the true-arm distributions. We endow $\mathcal{P}(\Re)$ with the topology of weak convergence, or equivalently, with the Lévy metric (see Section 2 for definitions). Another nuance in our analysis is that the empirical distributions may not lie in $\mathcal{L}$. This is handled by projecting them on to $\mathcal{L}$ under a suitable metric.

Our proposed algorithm is a plugin strategy that involves solving the lower bound problem using the empirical distributions as a proxy for the actual arm distributions. This can be computationally demanding especially as the underlying samples in the empirical distribution become large. To ease the numerical burden we propose modifications that require solving the lower bound only order $\log n$ times till stage n of the algorithm (where n samples are generated). This modification substantially reduces the computation burden. We show that it is optimal upto a constant (Appendix J).

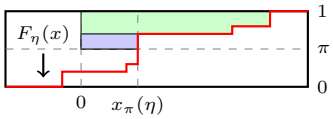
VaR problem: Our algorithm for CVaR, with $\mathrm{KL}^{\mathrm{L}}_{\inf}$ and $\mathrm{KL}^{\mathrm{U}}_{\inf}$ replaced by the corresponding functionals with the VaR constraints instead, is asymptotically optimal for this problem in complete generality (Section 4.2). Here, $\mathrm{KL}^{\mathrm{U}}_{\inf}$ and $\mathrm{KL}^{\mathrm{L}}_{\inf}$ have closed form representations. However, they are no longer jointly-continuous in the Lévy metric, which introduces new technical challenges in the analysis of the algorithm.

2 Background

For $K \geq 2$, let $\mathcal{M} = \mathcal{L}^K$ denote the collection of all K-vectors of distributions, $\nu = (\nu_1, \ldots, \nu_K)$, such that for all i, ν_i belongs to $\mathcal{L}$. Let $\mu \in \mathcal{M}$ be the given bandit problem, and $\pi \in (0,1)$ denote the fixed confidence level. For $\eta \in \mathcal{P}(\Re)$, let $F_\eta(y) = \eta((-\infty, y])$ denote the CDF function for η, and let $m(\eta)$ denote mean of measure η.

VaR, CVaR: With the above notation, VaR at level π for the distribution η, denoted as $x_\pi(\eta)$, equals $\min\{z \in \Re : F_\eta(z) \geq \pi\}$. Since $F_\eta(.)$ is a non-decreasing and right-continuous function, the minimum in the expression of VaR is always attained. Define CVaR at level π, $c_\pi(\eta)$, as

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



If η has a density in a neighbourhood around x_π, then $c_\pi(\eta) = \mathbb{E}_\eta\left(X | X \geq x_\pi(\eta)\right)$, i.e., it measures the average loss conditioned on the event that losses are larger than the VaR.

In the figure above, the total shaded area (green and blue regions, together) divided by $1-\pi$ denotes the CVaR of the measure whose CDF function is displayed in red. To see this, observe that the first term in the expression above, scaled by $(1-\pi)$, equals the blue region. The integral in the second term when simplified using integration by parts can be seen to equal the green region. There are alternative formulations of CVaR, which we state without proofs.

$$c_\pi(\eta) \;=\; \frac{1}{1-\pi}\int_{p\in[\pi,1]} x_p(\eta)dp \;=\; \min_{x_0 \in \Re}\left\{x_0 + \frac{1}{1-\pi}\mathbb{E}_\eta\left((X - x_0)_+\right)\right\} \tag{2}$$

$$= \max_{v \in M^+(\Re)} \frac{1}{1-\pi}\int_{\Re} y dv(y) \;\text{ s.t. }\; \forall y,\; dv(y) \leq d\eta(y) \;\text{ and }\; \int_{\Re} dv(y) = 1-\pi, \tag{3}$$

where $(x)_+$ denotes $\max\{0, x\}$ and $M^+(\Re)$ denotes collection of all non-negative measures on $\Re$.

From (2), since $c_\pi(\eta)$ is a minimum of linear functions of η, it is a concave function of η. Thus, the $\mathrm{KL}^{\mathrm{U}}_{\inf}$ problem in (1) is a convex optimization problem, while the $\mathrm{KL}^{\mathrm{L}}_{\inf}$ problem is not, since the $c_\pi(.)$ constraint makes the feasible region non-convex. See, [50] for a comprehensive tutorial on the two tail-risk measures, and their properties.

Parametric case: Using the definition of VaR, it can be argued that $x_\pi(\eta_\theta)$ is a monotonically increasing function of θ when η_θ belongs to a canonical SPEF with parameter θ, as is the mean. The first formulation in 2 then gives that $c_\pi(\eta_\theta)$ is also monotonically increasing. Thus, the problem of identifying the best-(CVaR/VaR/mean-CVaR) arm is equivalent to identifying that with minimum mean. See Appendix A for details.

General case: For η in class $\mathcal{L}$, the moment-constraint limits the minimum and maximum possible values of VaR and CVaR, as discussed in the following lemma (proof in Appendix B).

Lemma 2.1. *For* $\eta \in \mathcal{L}$, $c_\pi(\eta) \in D \triangleq [-B^{\frac{1}{1+\epsilon}}, B^{\frac{1}{1+\epsilon}}(1-\pi)^{\frac{-1}{1+\epsilon}}]$ *and* $x_\pi(\eta) \in C \triangleq [-B^{\frac{1}{1+\epsilon}}\pi^{\frac{-1}{1+\epsilon}}, B^{\frac{1}{1+\epsilon}}(1-\pi)^{\frac{-1}{1+\epsilon}}]$.

Topology of weak convergence and the Lévy metric: Let ϕ be a bounded and continuous function on $\Re$, $\delta > 0$, and $x \in \Re$. Consider the topology on $\mathcal{P}(\Re)$, generated by the base sets of the form $\mathcal{U}(\phi, x, \delta) = \left\{\eta \in \mathcal{P}(\Re) : |\int_{\Re}\phi(y)d\eta(y) - x| < \delta\right\}$. Weak convergence of sequence κ_n to κ, denoted as $\kappa_n \overset{D}{\Rightarrow} \kappa$, is convergence in this topology (see, [23, Section D.2]). It is equivalent to that in the Lévy metric on $\mathcal{P}(\Re)$, (denoted by d_L), defined next (see, [8, Theorem 6.8], [23, Theorem D.8]). For $\eta, \kappa \in \mathcal{P}(\Re)$, $d_L(\eta, \kappa)$ equals $\inf\{\delta > 0 : F_\eta(x-\delta) - \delta \leq F_\kappa(x) \leq F_\eta(x+\delta) + \delta,\ \forall x \in \Re\}$. Additionally, the metric space $(\mathcal{P}(\Re), d_L)$ is complete and separable.

3 Lower bound

We consider δ-correct algorithms for identifying the arm with minimum CVaR, acting on bandit problems in $\mathcal{M}$. While ensuring δ-correctness, the aim is to minimize the sample complexity, i.e., expected number of samples generated by the algorithm before it terminates. As is well known, the δ-correctness property imposes a lower bound on the sample complexity of such algorithms.

Let $\mu \in \mathcal{M}$ denote the given bandit problem. Henceforth, for ease of notation, we assume without loss of generality that the best-CVaR arm in μ is arm 1. Let Σ_K denote the probability simplex in $\Re^K$, $\mathcal{A}_j$ denote the collection of all bandit problems in $\mathcal{M}$ which have arm j as the best-CVaR arm, τ_δ be the stopping time for the δ-correct algorithm, $N_a(\tau)$ denote the number of times arm a has been sampled by the algorithm, and for a set S, let S^o denote its interior. It is easy to deduce using standard arguments (see, e.g., [40]) that for a δ-correct algorithm acting on $\mu \in \mathcal{A}_1$,

$$\mathbb{E}\left(\tau_\delta\right) \geq V(\mu)^{-1} \log \frac{1}{4\delta} \text{ where } V(\mu) = \sup_{t \in \Sigma_K} \inf_{\nu \in \mathcal{A}_1^c} \sum_{a=1}^{K} t_a \, \mathrm{KL}(\mu_a, \nu_a), \text{ and } \mathcal{A}_j^c = \mathcal{M} \setminus \mathcal{A}_j. \quad (4)$$

Lemma 3.1. *For $\mu \in \mathcal{A}_1$, the inner minimization problem in $V(\mu)$ equals $\min_{j \neq 1} \inf_{x \leq y} \left\{ t_1 \, \mathrm{KL}^{\mathrm{U}}_{\inf}(\mu_1, y) + t_j \, \mathrm{KL}^{\mathrm{L}}_{\inf}(\mu_j, x) \right\}$, and hence*

$$V(\mu) = \sup_{t \in \Sigma_K} \min_{j \neq 1} \inf_{x \leq y} \left\{ t_1 \, \mathrm{KL}^{\mathrm{U}}_{\inf}(\mu_1, y) + t_j \, \mathrm{KL}^{\mathrm{L}}_{\inf}(\mu_j, x) \right\}. \quad (5)$$

Remark 3.1. Without any restriction, i.e., if $\mathcal{L} = \mathcal{P}(\Re)$, then for CVaR-problem, for $y \in \Re$ and $\eta \in \mathcal{L}$, $\mathrm{KL}^{\mathrm{U}}_{\inf}(\eta, y) = 0$. This is essentially because η can be perturbed in KL only slightly by shifting an arbitrarily small mass from the lower tail to the extreme right, so that the CVaR constraint is satisfied. Thus, without any restrictions, $V(\mu) = 0$ (see, [2, Lemma 1, Theorem 3] for similar results in selecting the arm with the largest mean setting). However, we later solve the VaR-problem without any assumptions on the distributions, i.e., $\mathcal{L} = \mathcal{P}(\Re)$. The lower bound for the VaR problem is as in (4), with $\mathrm{KL}^{\mathrm{L}}_{\inf}$ and $\mathrm{KL}^{\mathrm{U}}_{\inf}$ in the representation in (5) defined with VaR constraints, instead.

A proof of Lemma 3.1 can be found in Appendix C.1. Let $t^* : \mathcal{M} \to 2^{\Sigma_K}$. In particular, for $\nu \in \mathcal{M}$, let $t^*(\nu)$ denote the set of maximizers in the $V(\nu)$ optimization problem in (5). A key nuance of our algorithm and the related analysis is that the empirical distribution may not belong to the class $\mathcal{M}$. The algorithm first projects the empirical distribution to class $\mathcal{L}$, then solves for the optimal t^* in (5) for the projected distributions, and samples the arms in proportion to the computed t^* (Section 4). For appropriate choices of the projection maps, the following lemmas guarantee that as the empirical distributions converge to the actual arm-distributions (in the weak topology), the t^* computed by the algorithm converge to the optimal weights corresponding to μ.

Lemma 3.2. *$\mathcal{L}$ is a compact set in the topology of weak convergence and is a uniformly integrable collection of random variables. When restricted to $\mathcal{L} \times D^o$, $\mathrm{KL}^{\mathrm{L}}_{\inf}$ and $\mathrm{KL}^{\mathrm{U}}_{\inf}$ are both jointly continuous functions of the arguments. Moreover, for fixed x, $\mathrm{KL}^{\mathrm{U}}_{\inf}(\nu, x)$ is a convex function of ν.*

Lemma 3.3. *t^* is an upper-hemicontinuous correspondence. For $\nu \in \mathcal{M}^o$, $t^*(\nu)$ is a convex set.*

In Lemma 3.2, we restrict to the interior of D as $\mathrm{KL}^{\mathrm{U}}_{\inf}(\cdot, B^{\frac{1}{1+\epsilon}}(1-\pi)^{\frac{-1}{1+\epsilon}})$ and $\mathrm{KL}^{\mathrm{L}}_{\inf}(\cdot, -B^{\frac{1}{1+\epsilon}})$ are not continuous (Remark C.2). In Lemma 3.3, we only need to eliminate distributions with these extreme CVaRs (there are only two such distributions. See Remark C.1). Lemma 3.3 above and Theorem 4.1 (optimality and δ-correctness of the proposed algorithm) hold for distributions with CVaR in D^o. For ease of notation, we restrict μ to lie in the interior of $\mathcal{M}$.

The proofs of the above two lemmas are technically challenging and involve nuanced analysis. Detailed steps are given in Appendix C.2 and C.3. We first prove joint lower- and upper-semicontinuity of the KL-projection functionals separately. These rely on various properties of the weak convergence of probability measures in $\mathcal{L}$, the dual representations for $\mathrm{KL}^{\mathrm{L}}_{\inf}$ and $\mathrm{KL}^{\mathrm{U}}_{\inf}$ (see Theorem 3.4), properties of CVaR for probability measures in $\mathcal{L}$, and the classical Berge's theorem (see, [52]) for continuity of the optimal value and the set of optimizers for a parametric optimization problem. We then use these to prove the continuity in Lemma 3.3. Convexity follows since t^* is the set of maximizers of a concave function over a convex, compact set.

Understanding the lower bound: Our proposed algorithm requires repeated evaluations of the lower bound in (4) at its estimates of μ. To facilitate this, we now provide a more tractable characterization

of the two KL-projection functionals, and in particular, (5). We also discuss the statistical and computational implications of these alternative characterizations.

For $\eta \in \mathcal{P}(\Re)$, let $\mathrm{Supp}(\eta)$ denote the collection of points in the support of measure η. For $v \in D^o$, $x_0 \in C$, $\boldsymbol{\lambda} \in \Re^3$, $\boldsymbol{\gamma} \in \Re^2$, and $X \in \Re$, set $g^U(X, \boldsymbol{\lambda}, v) = 1 + \lambda_1 v - \lambda_2(1-\pi) + \lambda_3(|X|^{1+\epsilon} - B) - (\lambda_1 X (1-\pi)^{-1} - \lambda_2)_+$, and $g^L(X, \boldsymbol{\gamma}, v, x_0) = 1 - \gamma_1(v - x_0 - (X - x_0)_+ (1-\pi)^{-1}) - \gamma_2(B - |X|^{1+\epsilon})$. Furthermore, define $\hat{S}(v) = \left\{\lambda_1 \geq 0, \lambda_2 \in \Re, \lambda_3 \geq 0 : \forall x \in \Re,\ g^U(x, \boldsymbol{\lambda}, v) \geq 0\right\}$, and $\mathcal{R}_2(x_0, v)$ to be $\left\{\gamma_1 \geq 0, \gamma_2 \geq 0 : \forall y \in \Re,\ g^L(y, \boldsymbol{\gamma}, v, x_0) \geq 0\right\}$. Notice that these are convex sets.

Theorem 3.4. *For $\eta \in \mathcal{P}(\Re)$ and $v \in D^o$,*

(a)

$$\mathrm{KL}^{\mathrm{U}}_{\inf}(\eta, v) = \max_{\boldsymbol{\lambda} \in \hat{S}(v)} \mathbb{E}_\eta \left(\log\left(g^U(X, \boldsymbol{\lambda}, v)\right)\right).$$

The maximum in this expression is attained at a unique point $\lambda^* \in \hat{S}(v)$. The unique probability measure $\kappa^* \in \mathcal{L}$ that achieves infimum in the primal problem satisfies $\frac{d\kappa^*}{d\eta}(y) = \left(g^U(y, \boldsymbol{\lambda}^*, v)\right)^{-1}$, for $y \in \mathrm{Supp}(\eta)$. Moreover, it has mass on at most 2 points outside $\mathrm{Supp}(\eta)$. Furthermore, for $y' \in \{\mathrm{Supp}(\kappa^*) \setminus \mathrm{Supp}(\eta)\}$, $g^U(y', \boldsymbol{\lambda}^*, v) = 0$.

(b)

$$\mathrm{KL}^{\mathrm{L}}_{\inf}(\eta, v) = \min_{x_0 \in [-\left(\frac{B}{\pi}\right)^{\frac{1}{1+\epsilon}}, v]} \max_{\boldsymbol{\gamma} \in \mathcal{R}_2(x_0, v)} \mathbb{E}_\eta \left(\log\left(g^L(X, \boldsymbol{\gamma}, v, x_0)\right)\right).$$

For a fixed x_0, the maximum in the inner problem is attained at a unique $\boldsymbol{\gamma}^*$ in $\mathcal{R}_2(x_0, v)$. The unique probability measure $\kappa^* \in \mathcal{L}$ achieving infimum in the primal problem satisfies $\frac{d\kappa^*}{d\eta}(y) = \left(g^L(y, \boldsymbol{\gamma}^*, v, x_0)\right)^{-1}$, for $y \in \mathrm{Supp}(\eta)$. Moreover, size of the set $\{\mathrm{Supp}(\kappa^*) \setminus \mathrm{Supp}(\eta)\}$ is at most 1, and for $y' \in \{\mathrm{Supp}(\kappa^*) \setminus \mathrm{Supp}(\eta)\}$, $g^L(y' \boldsymbol{\gamma}^*, v, x_0) = 0$.

These dual formulations help in reformulating the lower bound optimization problem in (5) as optimization over reals. A computationally more efficient approach for this is to consider the joint dual of the inner optimization problem in (5).

For $\eta_1, \eta_2 \in \mathcal{P}(\Re)$, and non-negative weights α_1, α_2, let

$$Z = \inf_{x \leq y} \left\{\alpha_1 \mathrm{KL}^{\mathrm{U}}_{\inf}(\eta_1, y) + \alpha_2 \mathrm{KL}^{\mathrm{L}}_{\inf}(\eta_2, x)\right\}. \tag{6}$$

For $y \in \Re$, $\boldsymbol{\lambda} \in \Re^2$, $\boldsymbol{\rho} \in \Re^2$, and $\boldsymbol{\gamma}_2 \in \Re$, let $h^L(y, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho}, x_0) = 1 - \lambda_1 + \gamma_2(|y|^{1+\epsilon} - B) + \rho_1(x_0 + (y - x_0)_+(1-\pi)^{-1})$, and $h^U(y, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho}) = 1 + \lambda_1 + \lambda_2(|y|^{1+\epsilon} - B) - (\rho_2 + (\rho_1 y - \rho_2)_+(1-\pi)^{-1})$. For $x_0 \in C$, define the convex region $\mathcal{D}_{x_0}$ to be collection of $\lambda_1 \in \Re$, $\rho_2 \in \Re$, $\lambda_2 \geq 0$, $\gamma_2 \geq 0$, and $\rho_1 \geq 0$, such that for all $y \in \Re$, $h^L(y, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho}, x_0) \geq 0$ and $h^U(y, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho}) \geq 0$.

Proposition 3.5. *For $\eta_1, \eta_2 \in \mathcal{P}(\Re)$ and weights $\alpha_1, \alpha_2 \in [0, 1]$, Z equals*

$$\begin{aligned}\min_{x_0 \in C} \max_{(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho}) \in \mathcal{D}_{x_0}} \quad & \alpha_1 \mathbb{E}_{\eta_1}\left(\log\left(h^U(X, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho})\right)\right) + \alpha_2 \mathbb{E}_{\eta_2}\left(\log\left(h^L(X, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho}, x_0)\right)\right) \\ & - \alpha_1 \log \alpha_1 - \alpha_2 \log \alpha_2 + (\alpha_1 + \alpha_2) \log(\alpha_1 + \alpha_2) - (\alpha_1 + \alpha_2) \log 2.\end{aligned}$$

An application of above to the empirical distributions ($\eta_a = \hat{\mu}_a(t)$) weighted by sample counts, i.e., $\alpha_a = N_a(t)$, for $a \in \{1, 2\}$, results in unweighted sums over the samples. Also observe that the representation in Proposition 3.5 is 1 dimension smaller compared to that obtained by using Theorem 3.4 in (6), and hence, is faster to optimize numerically.

Recall that $\mathrm{KL}^{\mathrm{U}}_{\inf}$ is a convex optimization problem, $\mathrm{KL}^{\mathrm{L}}_{\inf}$ is not (see Section 2). To handle this, we use the min formulation for CVaR in 2 to turn it into a one-dimensional family of linear constraints, which appears as the outer $\min_{x_0}$ in the expression in (b) in Theorem 3.4, and Proposition 3.5 above, with the range constraint from Lemma 2.1. This renders the remanining problem as a convex optimization problem. To simplify the CVaR constraint in $\mathrm{KL}^{\mathrm{U}}_{\inf}$, we use (3). Rest is the Lagrangian duality. Complete proofs for Theorem 3.4, and Proposition 3.5, are given in Appendix D.

The equality in Proposition 3.5, and the dual formulations in Theorem 3.4, are important statistically and computationally. First, our stopping rule will threshold the Z statistics to determine when to

safely stop. So we need to bound the deviations of Z. For this, we will use the dual formulations from Theorem 3.4 in (6) to construct mixtures of super-martingales that dominate the deviations of Z. Second, our sampling rule will sample according to the optimal proportions evaluated for the empirical distribution vector, $\hat{\mu}(t)$, in (5). For this, we use Proposition 3.5 in our experiments to solve the inner optimization problem in (5). These will be made precise in Section 4.

Computing a gradient for the objective of the maximisation problem (5), seen as a function of the sampling weights $\boldsymbol{t}$, takes one Z evaluation per suboptimal arm. The inner maximisation over $\mathcal{D}_{x_0}$ is a constrained concave program, for which standard algorithms apply. The outer $\min_{x_0}$ problem requires a different approach, as it is not even quasiconvex. Empirically it does become quasiconvex after seeing enough samples, so we employ a heuristic bisection search for which we measure the impact on the error probability (there is none). Numerical results are presented in Section 4.4.

4 The algorithm

Given a bandit problem $\mu \in \mathcal{M}$, our algorithm is a specification of three things: a sampling rule, a stopping rule, and a recommendation rule.

Sampling rule: At each iteration, the algorithm has access to the empirical distribution vector, $\hat{\mu}(t)$. It first projects $\hat{\mu}(t)$ to $\mathcal{L}^K$ using the projection map, Π, defined below. It then computes $t^*\left(\Pi\left(\hat{\mu}(t)\right)\right)$ and allocates samples using any reasonable tracking rule, like the C-tracking rule of [27], or the randomized tracking of [2]. The projection map $\Pi = (\tilde{\Pi}, \ldots, \tilde{\Pi})$, where $\tilde{\Pi} : \mathcal{P}(\Re) \to \mathcal{L}$, is given by $\tilde{\Pi}(\eta) \in \operatorname{argmin}_{\kappa \in \mathcal{L}} \sup_{x \in \Re} |F_\eta(x) - F_\kappa(x)|$. We show in the Appendix F that this projection has a simple form and can be computed easily.

Stopping rule: We use a modification of the generalized likelihood ratio test (GLRT) (see, [16]) as our stopping criteria. At any time, the empirical distribution vector, $\hat{\mu}(n)$, suggests an arm with minimum CVaR (empirically best-CVaR arm), say arm i. This is our null hypothesis, which we test against all the alternatives. Formally, the log of the GLRT statistic, denoted by $S_i(n)$, is $\inf_{\nu' \in \mathcal{A}_i^c} \sum_{a=1}^K N_a(n)\,\mathrm{KL}(\hat{\mu}_a(n), \nu'_a)$. This is exactly the scaled inner optimization problem in the expression of $V(\hat{\mu}(n))$ in (4), except that $\hat{\mu}(n)$ may not belong to $\mathcal{M}$ (recall the Z statistic defined in (6)). Let $Z_i(n)$ equal $\min_{a \neq i} \inf_{x \leq y} \left\{N_i(n)\,\mathrm{KL}^{\mathrm{U}}_{\inf}(\hat{\mu}_i(n), y) + N_a(n)\,\mathrm{KL}^{\mathrm{L}}_{\inf}(\hat{\mu}_a(n), x)\right\}$. It equals $S_i(n)$ when $\hat{\mu}(n) \in \mathcal{M}$ (Lemma 3.1). Our stopping rule corresponds to checking

$$Z_i(n) \geq \beta(n, \delta) \quad \text{where} \quad \beta(n, \delta) = \log\left((K-1)\,\delta^{-1}\right) + 5\log(n+1) + 2. \tag{7}$$

Recommendation rule: On stopping, the algorithm outputs the empirically-minimal CVaR arm.

4.1 Theoretical guarantees

For a given δ, let τ_δ denote the stopping time for the algorithm. The algorithm makes an error if at time τ_δ, there is an arm $j \neq 1$ such that $c_\pi(\hat{\mu}_j(t)) < c_\pi(\hat{\mu}_1(t))$. Let the error event be denoted by $\mathcal{E}$.

Theorem 4.1. *For $\delta > 0$ and $\mu \in \mathcal{M}^o$, the proposed algorithm with $\beta(t, \delta)$ chosen as in (7), satisfies $\mathbb{P}(\mathcal{E}) \leq \delta$, and $\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu(\tau_\delta)}{\log(1/\delta)} \leq \frac{1}{V(\mu)}$.*

We first give a proof sketch for δ-correctness part of the theorem. Proof ideas for sample complexity are presented later in this section. The detailed proof for Theorem 4.1 can be found in Appendix E.

δ-correctness: Recall that the algorithm makes an error if at time τ_δ, the empirically best-CVaR arm is not arm 1. As in the best-mean arm case, it can be argued that this probability is at most

$$\sum_{i=2}^{K} \mathbb{P}\left(\exists n : \; N_i(n)\,\mathrm{KL}^{\mathrm{U}}_{\inf}(\hat{\mu}_i(n), c_\pi(\mu_i)) + N_1(n)\,\mathrm{KL}^{\mathrm{L}}_{\inf}(\hat{\mu}_1(n), c_\pi(\mu_1)) \geq \beta\right). \tag{8}$$

The following proposition will be helpful in bounding each of the summands above. Setting $j = 1$, and $x = \log \frac{K-1}{\delta}$ in Proposition 4.2, along with β from (7), we get that each summand in (8) is at most $\delta/(K-1)$, proving that the proposed algorithm is δ-correct.

Proposition 4.2. *For $i \in [K], j \in [K]$, $i \neq j$, $h(n) = 5\log(n+1) + 2$, and $x \geq 0$,*

$$\mathbb{P}\left(\exists n : N_i(n)\,\mathrm{KL}^{\mathrm{U}}_{\inf}(\hat{\mu}_i(n), c_\pi(\mu_i)) + N_j(n)\,\mathrm{KL}^{\mathrm{L}}_{\inf}(\hat{\mu}_j(n), c_\pi(\mu_j)) - h(n) \geq x\right) \leq e^{-x}.$$

A key step in proving Proposition 4.2 is constructing mixtures of super-martingales that dominate the exponentials of $N_i(n)\,\mathrm{KL}^{\mathrm{U}}_{\inf}(\hat{\mu}_i(n), c_\pi(\mu_i))$ and $N_i(n)\,\mathrm{KL}^{\mathrm{L}}_{\inf}(\hat{\mu}_i(n), c_\pi(\mu_i))$. From Theorem 3.4(a) and (2), it can be shown that for fixed dual-variables, the objective is a sum of logs of random variables with mean at-most 1. Hence, its exponential is a non-negative candidate super-martingale. Since we want to bound the maximum over the dual parameters, we construct a mixture of these candidates, over the dual-parameters, and show that it dominates the exponential of $N_i(n)\,\mathrm{KL}^{\mathrm{U}}_{\inf}(\hat{\mu}_i(n), c_\pi(\mu_i))$.

Sample complexity: Our sample complexity proof follows that of [27] for a parametric family. However, we work with a more general non-parametric class, in which we establish continuity of the KL-projection functionals (Lemma 3.2). Our proof also differs from that in [2] in that we only have upper-hemicontinuity of the set of optimal sampling allocation (t^*) (Lemma 3.3). A nuance in our analysis is that the empirical distribution may not belong to the class $\mathcal{L}$, in which case we project the empirical distribution onto the class, and the sampling rule uses this projected distribution to compute t^*. Careful choice of the projection map aids in the proof of this result.

Computational complexity: The computational cost of these KL-projection functionals, and hence, that of the oracle weights, is linear in the number of samples taken (see, [2, 15, 30]). As a result, the overall run-time is quadratic in τ_δ. We propose a modification in which we update the weights only at the beginning of geometrically spaced times. This modification improves the computational-cost to almost linear in τ_δ, while its sample complexity is optimal up to a multiplicative constant depending on the choice of geometrical-spacing factor, thus providing a controlled trade-off between the two costs. We refer the reader to Appendix J for details of the algorithm and proofs for its theoretical guarantees.

Mean-CVaR problem: We now extend the methodology for the CVaR problem to the more general mean-CVaR problem. For a distribution $\eta \in \mathcal{L}$ (for example, a random loss in a financial investment), the metric associated with the "badness" of a distribution is $\alpha_1 m(\eta) + \alpha_2 c_\pi(\eta)$, for $\alpha_1 > 0$ and $\alpha_2 > 0$, and the best-arm is the one with minimum value of this conic combination of mean and CVaR. For $\alpha_1 = 0$ this is the CVaR-problem, which we have studied in this work. For $\alpha_2 = 0$, this is the mean-problem, extensively studied in [2, 27]. As in (1), we can define corresponding KL-projection functionals, with the CVaR constraints replaced with those on the modified metric. The above theory, with this updated $\mathrm{KL}^{\mathrm{L}}_{\inf}$ and $\mathrm{KL}^{\mathrm{U}}_{\inf}$, gives the corresponding results for this setting. In particular, the lower bound on $\mathbb{E}(\tau_\delta)$ for δ-correct algorithms for mean-CVaR BAI is given by $V(\mu)^{-1} \log \frac{1}{4\delta}$, where $V(\mu)$ is defined in (5) with the updated $\mathrm{KL}^{\mathrm{U}}_{\inf}$ and $\mathrm{KL}^{\mathrm{L}}_{\inf}$.

Theorem 4.3 (Informal)**.** *For $\mu \in \mathcal{M}^o$, the proposed algorithm for CVaR, with $\mathrm{KL}^{\mathrm{U}}_{\inf}$ and $\mathrm{KL}^{\mathrm{L}}_{\inf}$ defined with mean-CVaR constraints instead, is δ-correct and asymptotically optimal.*

The proof of this theorem parallels that for CVaR above. We give the formal statement with proof-details in Appendix H.

4.2 The VaR problem

In this section we present the main ideas for an analogous approach for the optimum VaR-problem. Here, we will not impose any conditions on the arm-distributions, as the VaR lower bound is defined without it, i.e., $\mathcal{L} = \mathcal{P}(\Re)$. For a probability measure η, let $F_\eta(y)$, denote its CDF evaluated at y and $F^-_\eta(y) = \lim_{z \uparrow y} F_\eta(z)$ denote the left limit of the CDF. Moreover, for $r, q \in (0,1)$ let $d_2(r,q)$ denote the KL divergence between the Bernoulli random variables with mean r and q. For $y \in \Re$, $\mathrm{KL}^{\mathrm{L}}_{\inf}(\eta, y)$ and $\mathrm{KL}^{\mathrm{U}}_{\inf}(\eta, y)$ be defined as in (1), with VaR constraints, instead.

Lemma 4.4. $\mathrm{KL}^{\mathrm{L}}_{\inf}(\eta, y) = d_2(\min\{F_\eta(y), \pi\}, \pi)$ *and* $\mathrm{KL}^{\mathrm{U}}_{\inf}(\eta, y) = d_2(\max\{F^-_\eta(y), \pi\}, \pi)$.

Unlike in the CVaR-problem, we show that $\mathrm{KL}^{\mathrm{L}}_{\inf}$ and $\mathrm{KL}^{\mathrm{U}}_{\inf}$ for the VaR problem are not jointly continuous functionals (see Remark G.2). The discontinuity occurs at y being the jump points of F_η in Lemma 4.4 above. However, we prove in Appendix G (Corollary G.3.1) that the set of optimal proportions, t^*, is still upper-hemicontinuous and convex.

The algorithm for CVaR with $\mathrm{KL}^{\mathrm{U}}_{\inf}$ and $\mathrm{KL}^{\mathrm{L}}_{\inf}$ replaced by those in the lemma above, and setting $\beta(t,\delta) = 6\log\left(1 + \log\frac{t}{2}\right) + \log\frac{K-1}{\delta} + 8\log\left(1 + \log\frac{K-1}{\delta}\right)$, we get an algorithm for VaR-problem.

Theorem 4.5 (Informal)**.** *The proposed algorithm for the VaR-problem is δ-correct and asymptotically optimal.*

We refer the reader to Appendix G for a detailed discussion of the VaR-problem and proofs.

4.3 Tight $\mathrm{KL}_{\mathrm{inf}}$-based confidence intervals for CVaR

We now present tight anytime-valid confidence interval for CVaR of distributions in $\mathcal{L}$. Let $\hat{\eta}_n$ denote the empirical distribution corresponding to n samples from $\eta \in \mathcal{L}$. Our proposed upper (U_n) and lower (L_n) confidence intervals for $c_\pi(\eta)$ are of the form $U_n = \max\left\{x \in \Re : \, n\,\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}}(\hat{\eta}_n, x) \leq C\right\}$ and $L_n = \min\left\{x \in \Re : \, n\,\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}(\hat{\eta}_n, x) \leq C\right\}$, for appropriate threshold, $C \approx \log \delta^{-1} + 3 \log n$. Let $\hat{x}_{\pi,n}$ denote the πth quantile for $\hat{\eta}$. Recall that the popular truncation based estimator for $c_\pi(\eta)$ is given by $\hat{c}_{\pi,n} = n^{-1}(1-\pi)^{-1} \sum_i X_i \mathbb{1}(\hat{x}_{\pi,n} \leq X_i \leq u_n)$, for appropriately chosen truncation levels, u_n (see, [39]). Observe that there are 2 sources of error in this estimator, first, the estimation of the quantile, and second, the estimation of the tail-expectation. On the other hand, our confidence intervals do not rely on estimation of the true quantile, x_π. In Appendix I, we show that even given the correct estimation of $\hat{x}_{\pi,n}$, classical confidence intervals for $\hat{c}_{\pi,n}$ are typically wider than those based on $\mathrm{KL}^{\mathrm{U}}_{\mathrm{inf}}$ and $\mathrm{KL}^{\mathrm{L}}_{\mathrm{inf}}$.

4.4 Numerical Results

This is only a brief teaser section on the experiments, which are detailed in Appendix K. We are interested in the question whether the asymptotic sample complexity result of Theorem 4.1 is representative at reasonable confidence levels δ. Whether this is the case or not differs greatly between pure exploration setups. [27] see state-of-the-art numerical results in Bernoulli arms for Track-and-Stop with $\delta = 0.1$, while [22] present a Minimum Threshold problem instance where the Track-and-Stop asymptotics have not kicked in yet at $\delta = 10^{-20}$. Our experiments confirm that our approach is indeed practical at moderate confidence δ.

In our experiments we implement a version of Track-and-Stop including C-tracking and forced exploration and apply it to Fisher-Tippett ($F(\mu, \sigma, \gamma)$), Pareto ($P(\mu, \sigma, \gamma)$), and mixtures of Fisher-Tippett arms (these heavy tailed distributions arise in extreme value theory).

Figure 1 shows the distribution of the stopping time as a function of δ in an experiment with three arms: arm 1 is a uniform mixture of $F(-1, 0.5, 0.4)$ and $F(-3, 0.5, -0.4)$, arm 2 is $P(0, 0.2, 0.55)$ and arm 3 is $F(-0.5, 1, 0.1)$ with respective CVaRs at quantile 0.6 being -0.1428, 0.974 and 1.547. We select $\epsilon = 0.7$ and $B = 4.5$. This is a moderately hard problem of complexity $V^{-1}(\mu)^* = 49.7$. We conclude that even at moderate δ the average sample complexity closely matches the lower bound, especially after adjusting it for the lower-order terms in the employed stopping threshold $\beta(n, \delta)$. This demonstrates that our asymptotic optimality is in fact indicative of the performance in practice.

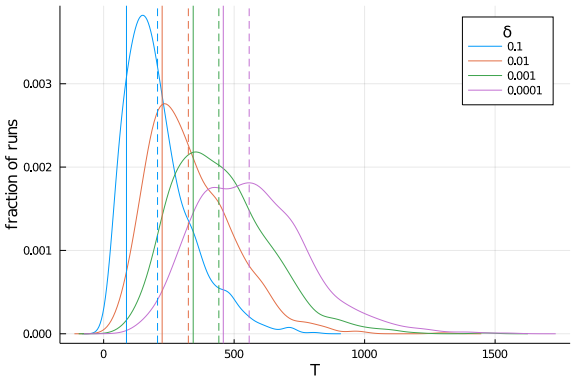


Figure 1: Histogram of stopping times based among 1000 runs on 3 Fisher-Tippett arms, as a function of confidence δ. Vertical bars indicate the lower bound (4) (solid), and a version adjusted to our stopping threshold (7) (dashed).

We do additional experiments to show the dependence of performance of the algorithm on number of arms and input parameter, B. We see that the average stopping time of our algorithm increases linearly in number of arms. Moreover, the sample complexity is also sensitive to B, indicating importance of correctly estimating it. We refer the reader to Appendix K for details of these experiments.

Conclusion: We developed asymptotically optimal algorithms that identify the arm with the minimum risk, measured in terms of CVaR, VaR, or a conic combination of mean and CVaR. Our algorithms operate in non-parametric settings with possibly heavy-tailed distributions. Although similar plug-and-play algorithms have been developed in simpler settings, our algorithms for tail-risk measures require more nuanced analysis. The techniques developed may be generalizable to a much broader class of problems.

References

[1] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR, 2013.

[2] 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*, volume 117 of *Proceedings of Machine Learning Research*, pages 61–110. PMLR, 08 Feb–11 Feb 2020.

[3] E. Arıkan and J. Fichtinger. The risk-averse newsvendor problem under spectral risk measures: A classification with extensions. *European Journal of Operational Research*, 256(1):116–125, 2017.

[4] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[5] J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, May 2002.

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

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

[9] A. Blum and A. Kalai. Universal portfolios with and without transaction costs. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 309–313. ACM, 1997.

[10] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] S. Bubeck, N. 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.

[12] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tails. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[13] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theor. Comput. Sci.*, 412(19):1832–1852, 2011.

[14] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122 – 142, 1996.

[15] 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*, 41(3):1516–1541, 2013.

[16] H. Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

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

[18] L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[19] V. H. de la Pena, M. J. Klass, and T. L. Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933, 2004.

[20] R. Degenne and W. M. Koolen. Pure exploration with multiple correct answers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14591–14600. Curran Associates, Inc., 2019.

[21] R. Degenne and W. M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems 32*, pages 14564–14573, 2019.

[22] R. Degenne, W. M. Koolen, and P. Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems 32*, pages 14465–14474, 2019.

[23] A. Dembo and O. Zeitouni. Large deviations techniques and applications. corrected reprint of the second (1998) edition. stochastic modelling and applied probability, 38, 2010.

[24] E. Even-Dar, M. Kearns, and J. Wortman. Risk-sensitive online learning. In J. L. Balcázar, P. M. Long, and F. Stephan, editors, *Algorithmic Learning Theory*, pages 199–213, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[25] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

[26] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

[27] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

[28] P. W. Glynn, A. Zeevi, et al. Bounding stationary expectations of markov processes. In *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*, pages 195–214. Institute of Mathematical Statistics, 2008.

[29] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.

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

[31] L. J. Hong, Z. Hu, and G. Liu. Monte carlo methods for value-at-risk and conditional value-at-risk: a review. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 24(4):1–37, 2014.

[32] S. R. Howard and A. Ramdas. Sequential estimation of quantiles with applications to A/B-testing and best-arm identification. *arXiv preprint arXiv:1906.09712*, 2019.

[33] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

[34] S. Juneja and S. Krishnasamy. Sample complexity of partition identification using multi-armed bandits. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1824–1852. PMLR, 25–28 Jun 2019.

[35] 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*, pages 11272–11281. Curran Associates, Inc., 2019.

[36] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[37] 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*, 17(1):1–42, 2016.

[38] E. Kaufmann and W. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv preprint arXiv:1811.11419*, 2018.

[39] P. L.A., K. Jagannathan, and R. Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In H. D. III and A. 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.

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

[41] D. Luenberger. *Optimization by Vector Space Methods*. Series in Decision and Control. Wiley, 1969.

[42] O.-A. Maillard. Robust risk-averse stochastic multi-armed bandits. In S. Jain, R. Munos, F. Stephan, and T. Zeugmann, editors, *Algorithmic Learning Theory*, pages 218–233, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[43] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

[44] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

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

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

[47] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[48] R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[49] A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[50] 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*, pages 270–294. Informs, 2008.

[51] T. Sawik. Optimization of cost and service level in the presence of supply chain disruption risks: Single vs. multiple sourcing. *Computers amd Operations Research*, 51:11 – 20, 2014.

[52] R. K. Sundaram. *A First Course in Optimization Theory*. Cambridge University Press, 1996.

[53] W. 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.

[54] H. Victor et al. A general class of exponential inequalities for martingales and ratios. *Annals of probability*, 27(1):537–564, 1999.

[55] M. K. Warmuth and D. Kuzmin. Online variance minimization. In G. Lugosi and H. U. Simon, editors, *Learning Theory*, pages 514–528, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[56] D. Williams. *Probability with martingales*. Cambridge university press, 1991.

[57] R. C. Williamson and A. K. Menon. Fairness risk measures. *arXiv preprint arXiv:1901.08665*, 2019.

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , See Section 1.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]