Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Survey Track

A Survey of Risk-Aware Multi-Armed Bandits

Vincent Y. F. Tan[1] , **Prashanth L.A.**[2] , **Krishna Jagannathan**[2]
[1]National University of Singapore
[2]IIT Madras
vtan@nus.edu.sg, prashla@cse.iitm.ac.in, krishnaj@ee.iitm.ac.in

Abstract

In several applications such as clinical trials and financial portfolio optimization, the expected value (or the average reward) does not satisfactorily capture the merits of a drug or a portfolio. In such applications, *risk* plays a crucial role, and a risk-aware performance measure is preferable, so as to capture losses in the case of adverse events. This survey aims to consolidate and summarise the existing research on risk measures, specifically in the context of multi-armed bandits. We review various risk measures of interest, and comment on their properties. Next, we review existing concentration inequalities for various risk measures. Then, we proceed to defining risk-aware bandit problems, We consider algorithms for the regret minimization setting, where the exploration-exploitation trade-off manifests, as well as the best-arm identification setting, which is a pure exploration problem—both in the context of risk-sensitive measures. We conclude by commenting on persisting challenges and fertile areas for future research.

1 Introduction

The multi-armed bandit (MAB) problem studies the problem of online learning with partial feedback that exemplifies the exploration-exploitation tradeoff. This problem has a long history dating back to Thompson [1933] and finds a wide variety of applications from clinical trials to financial portfolio optimization. In the stochastic MAB problem, a player chooses or pulls one among several arms, each defined by a certain reward distribution. The player wishes to maximize his reward or find the best arm in the face of the uncertain environment the distributions are *a priori* unknown. There are two general sub-problems in the MAB literature, namely, regret minimization and best-arm identification (also called pure exploration). In the former, in which the exploration-exploitation trade-off manifests, the player wants to maximize his reward over a fixed time period. In the latter, the player simply wants to learn which arm is the best in either the quickest time possible with a given probability of success (the fixed-confidence setting) or he wants to do so with the highest probability of success given a fixed playing horizon (the fixed budget setting). In most of the MAB literature (see Lattimore and Szepesvari [2020] for an up-to-date survey), the metric of interest is defined simply as the mean of the reward distribution associated with the arm pulled.

However, in real-world applications, the mean does not satisfactorily capture the merits of certain actions. In such applications, *risk* plays an important role, and a risk-aware performance measure is often preferred over the average reward. For example, if we have an important face-to-face meeting downtown, we might want to choose a route that has a slightly higher expected travel time, but whose variance is small. Some risk measures include the mean-variance [Markowitz, 1952], the conditional value-at-risk [Artzner *et al.*, 1999; Rockafellar and Uryasev, 2000], spectral risk measures [Acerbi, 2002], and cumulative prospect theory [Tversky and Kahneman, 1992]. This short survey aims to consolidate these various risk measures and to comment on their concentration properties. These properties are then used to describe the latest algorithmic developments for risk-aware bandits in regret minimization as well as best arm identification settings.

2 Preliminaries

For forming estimators of risk measures, we require the empirical distribution function (EDF), which we define below.

Definition 1. *Given i.i.d. samples $\{X_i\}_{i=1}^n$ from the distribution F of a random variable (r.v.) X, the EDF is*

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \le x\} \quad \textit{for any} \quad x \in \mathbb{R}.$$

To establish concentration bounds for risk measures, one requires an assumption that bounds the tail of the distribution. In this article, we consider sub-Gaussian distributions.

Definition 2. *A r.v. X is* sub-Gaussian with parameter σ^2 *if its cumulant generating function*

$$\log \mathbb{E}[\exp(rX)] \le r^2\sigma^2/2 \quad \textit{for any} \quad r \in \mathbb{R}.$$

See Theorem 2.1 of Wainwright [2019] for equivalent characterizations. A sub-Gaussian r.v. X satisfies the following bound for any $\epsilon > 0$:

$$\mathbb{P}\left(X \ge \epsilon\right) \le \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right),$$

For unifying risk measures, we require the notion of the Wasserstein distance, which is defined below.

Definition 3. *The* Wasserstein distance *between two cumulative distribution functions (CDFs) F_1 and F_2 on $\mathbb{R}$ is*

$$W_1(F_1, F_2) := \inf_{F \in \Gamma(F_1, F_2)} \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}F(x, y),$$

where $\Gamma(F_1, F_2)$ is the set of couplings of F_1 and F_2.

The Wasserstein distance between two CDFs. F_1 and F_2 of two r.v.s X and Y, respectively, may be alternatively written as follows:

$$W_1(F_1, F_2) = \sup |\mathbb{E}(f(X) - \mathbb{E}(f(Y))|$$
$$= \int_{-\infty}^{\infty} |F_1(s) - F_2(s)| \mathrm{d}s = \int_0^1 |F_1^{-1}(\beta) - F_2^{-1}(\beta)| \mathrm{d}\beta,$$

where the supremum is over all functions $f : \mathbb{R} \to \mathbb{R}$ that are 1-Lipschitz.

3 Estimation and Concentration of Risk Measures

3.1 Mean-Variance

We present a measure that effectively balances between expected reward and variability. Although there are a large number of models that resolve the tension between return and risk—such as the *Sharpe ratio* [Sharpe, 1966] and the *Knightian uncertainty* [Knight, 1921], we commence our discussion on one of the most popular measures, namely the mean-variance proposed by Markowitz [1952].

Definition 4. *The mean-variance of a r.v. X with mean μ and variance σ^2 and risk tolerance γ is* $\mathrm{MV} = \gamma\mu - \sigma^2$.

We can recover two extreme cases by considering the extremal values of the risk tolerance γ. When $\gamma \to \infty$, maximizing MV simply reduces to maximizing the mean, turning the mean-variance MAB problem into a standard reward maximizing problem. When $\gamma = 0$, the mean-variance MAB problem reduces to a variance-minimization problem. When one only has access to samples, one can estimate the mean-variance by using plug-in estimates for the mean and variance. Concentration results for the mean-variance can be found in Zhu and Tan [2020] and Sani *et al.* [2012].

3.2 Lipschitz-Continuous Risk Measures

In Cassel *et al.* [2018], the authors consider general risk measures that satisfy a Lipschitz requirement under some norm in the space of distributions. Prashanth and Bhat [2020] use the notion of Wasserstein distance as the underlying norm in defining a continuous class of risk measures as given below.

Definition 5. *Let $(\mathcal{L}, W_1)$ denote the metric space of distributions. A risk measure $\rho(\cdot)$ is Lipschitz-continuous on $(\mathcal{L}, W)$ if there exists $L > 0$ such that, for any two distributions (CDFs) $F, G \in \mathcal{L}$, the following holds:*

$$|\rho(F) - \rho(G)| \leq L \, W_1(F, G). \tag{1}$$

Using the EDF F_n computed from n i.i.d. samples, we estimate the risk measure $\rho(F)$ satisfying (1) as follows: $\rho_n := \rho(F_n)$.

Theorem 1. *Let X be a sub-Gaussian r.v. with parameter σ^2. Suppose $\rho : \mathcal{L} \to \mathbb{R}$ is a Lipschitz risk measure with parameter L. Then, for every ϵ satisfying $\frac{256\sqrt{2}\sigma}{\sqrt{n}} < \frac{\epsilon}{L} < \frac{256\sqrt{2}\sigma}{\sqrt{n}} + 16\sigma\sqrt{2\mathsf{e}}$, we have*

$$\mathbb{P}\left(|\rho_n - \rho(X)| > \epsilon\right) \leq \exp\left(-\frac{n}{256\sigma^2\mathsf{e}}\Big(\frac{\epsilon}{L} - \frac{256\sqrt{2}\sigma}{\sqrt{n}}\Big)^2\right),$$

where e is Euler's number.

Conditional Value-at-Risk [Rockafellar and Uryasev, 2000], and the spectral risk measure [Acerbi, 2002] are two popular risk measures that are Lipschitz continuous. We describe these risk measures below. For other examples of risk measures satisfying (1), the reader is referred to Cassel *et al.* [2018] and Prashanth and Bhat [2020].

3.3 Conditional Value-at-Risk (CVaR)

Definition 6. *Let $(y)^+ := \max(y, 0)$. The CVaR at level $\alpha \in (0, 1)$ for a r.v. X is defined by*

$$\mathrm{CVaR}_\alpha(X) := \inf_{\xi \in \mathbb{R}} \left\{ \xi + \frac{1}{(1-\alpha)} \mathbb{E}\big[(X - \xi)^+ \big] \right\}.$$

It is well known [Rockafellar and Uryasev, 2000] that the infimum in the definition of CVaR above is achieved for $\xi = \mathrm{VaR}_\alpha(X)$, where $\mathrm{VaR}_\alpha(X) = \inf\{\xi : \mathbb{P}(X \leq \xi) \geq \alpha\}$ is the Value-at-Risk (VaR) of X at confidence level α. In financial applications, one prefers CVaR over VaR, since CVaR is coherent, while VaR is not [Artzner *et al.*, 1999]. CVaR also admits the following equivalent representation:

$$\mathrm{CVaR}_\alpha(X) = \mathrm{VaR}_\alpha(X) + \frac{1}{1-\alpha} \mathbb{E}\left[X - \mathrm{Var}_\alpha(X)\right]^+.$$

The following lemma shows that CVaR is a Lipschitz risk measure in the sense of Definition 5.

Lemma 1. *Suppose X and Y are r.v.s with CDFs F_X and F_Y, respectively, and $\alpha \in (0, 1)$. Then*

$$|\mathrm{CVaR}_\alpha(X) - \mathrm{CVaR}_\alpha(Y)| \leq \frac{1}{1-\alpha} W_1(F_X, F_Y).$$

Let $X_1, \ldots, X_n$ denote i.i.d. samples drawn from the distribution of X. $\mathrm{CVaR}_\alpha(X)$ can be estimated as follows:

$$c_{n,\alpha} = \inf_{\xi \in \mathbb{R}} \left\{ \xi + \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} (X_i - \xi)^+ \right\}.$$

Let Y denote a r.v. with distribution F_n. Then, it can be shown that

$$c_{n,\alpha} = \mathrm{CVaR}_\alpha(Y) = \hat{v}_{n,\alpha} + \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} (X_i - \hat{v}_{n,\alpha})^+,$$

where $\hat{v}_{n,\alpha} = \inf\{x \in \mathbb{R} : F_n(x) \geq \alpha\}$.

A concentration bound for CVaR estimation under a sub-Gaussianity assumption follows as a corollary to Theorem 1. Other works on CVaR estimation can be found in Brown [2007], Wang and Gao [2010], Kolla *et al.* [2019], Thomas and Learned-Miller [2019], Kagrecha *et al.* [2019], and Prashanth *et al.* [2020].

3.4 Spectral Risk Measures (SRMs)

Given a risk spectrum $\phi : [0,1] \to [0,\infty)$, the SRM M_ϕ associated with ϕ is defined by Acerbi [2002]

$$M_\phi(X) = \int_0^1 \phi(\beta) F_X^{-1}(\beta)\, \mathrm{d}\beta.$$

M_ϕ is a coherent risk measure if the risk spectrum is increasing and integrates to 1. Further, M_ϕ generalizes CVaR, since setting $\phi(\beta) = (1-\alpha)^{-1}\mathbb{I}\{\beta \geq \alpha\}$ leads to $M_\phi = \mathrm{CVaR}_\alpha(X)$.

The result below shows that an SRM is a Lipschitz risk measure if the risk spectrum ϕ is bounded.

Lemma 2. *Suppose $\phi(u) \leq K$ for all $u \in [0,1]$, and let X and Y be r.v.s with CDFs F_X and F_Y, respectively. Then*

$$|M_\phi(X) - M_\phi(Y)| \leq K\, W_1(F_X, F_Y).$$

Specializing the estimator $\rho_n = \rho(F_n)$ for SRM leads to the following estimator:

$$m_{n,\phi} = \int_0^1 \phi(\beta) F_n^{-1}(\beta)\, \mathrm{d}\beta.$$

Using Lemma 2 and Theorem 1, a concentration bound for SRM estimation can be derived in a straightforward manner.

Since CVaR and spectral risk measures are weighted averages of the underlying distribution quantiles, a natural alternative to a Wasserstein distance-based approach is to employ concentration results for quantiles; cf. Prashanth *et al.* [2020] and Pandey *et al.* [2021]. While such an approach can provide bounds with better constants, the resulting bounds also involve distribution-dependent quantities. In this survey, we have chosen the Wasserstein distance-based approach since it provides a unified bound for an abstract risk measure that is Lipschitz, and one can easily specialize this bound to handle CVaR, SRM and other risk measures. Secondly, a template confidence-bound type bandit algorithm can be arrived at using the unified concentration bound in Theorem 1.

3.5 Risk Measures based on Cumulative Prospect Theory (CPT)

We present a risk measure based on CPT, and this risk measure does not satisfy the Lipschitz condition in Definition 5. The CPT-value of an r.v. X is defined as [Prashanth *et al.*, 2016]

$$\mathcal{C}(X) := \int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) \mathrm{d}z - \int_0^\infty w^-\left(\mathbb{P}\left(u^-(X) > z\right)\right) \mathrm{d}z,$$

where $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$ are the utility functions that are assumed to be continuous, with $u^+(x) = 0$ when $x \leq 0$ and increasing otherwise, and with $u^-(x) = 0$ when $x \geq 0$ and decreasing otherwise, $w^+, w^- : [0,1] \to [0,1]$ are weight functions assumed to be continuous, non-decreasing and satisfy $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$.

Given a set of i.i.d. samples $\{X_i\}_{i=1}^n$, we form the EDFs $F_n^+(x)$ and $F_n^-(x)$ of the r.v.s $u^+(X)$ and $u^-(X)$, respectively, Using these EDFs, we estimate CPT-value as follows:

$$\overline{\mathcal{C}}_n = \int_0^\infty w^+\left(1 - F_n^+(x)\right) \mathrm{d}x - \int_0^\infty w^-\left(1 - F_n^-(x)\right) \mathrm{d}x,$$

The integrals above can be computed using the order statistics; see Jie *et al.* [2018].

We now present a concentration result for CPT-value estimation assuming bounded support for the underlying distribution; see Prashanth *et al.* [2016] and Jie *et al.* [2018] for the proof.

(A1) The weight functions $w^\pm$ are Hölder continuous with common order α and constant H, i.e.,

$$\sup_{x,y\in[0,1]: x\neq y} \frac{|w^\pm(x) - w^\pm(y)|}{|x-y|^\alpha} \leq H;$$

(A2) The utility functions $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$ are bounded above by M.

Proposition 1. *Under (A1) and (A2), for all $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\overline{\mathcal{C}}_n - \mathcal{C}(X)\right| \geq \epsilon\right) \leq 2\exp\left(-2n\left(\frac{\epsilon}{HM}\right)^{\frac{2}{\alpha}}\right).$$

For the sub-Gaussian case, one can use a truncated CPT-value estimator, and establish an exponentially decaying tail bound; see Bhat and Prashanth [2019] for details.

4 Regret Minimization

A *bandit instance* ν is a collection of K distributions $(\nu_1, \nu_2, \ldots, \nu_K)$. Each distribution ν_i has mean $\mu_i \in \mathbb{R}$ and without loss of generality, one may assume that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$. The agent interacts with the bandit environment at each time by pulling an arm $A_t \in \{1, \ldots, K\}$ and observing a reward X_{t,A_t} drawn from ν_{A_t}. The decision of which arm to pull depends on the previous arm pulls as well as rewards $\mathcal{H}_{t-1} = (A_1, X_{1,A_1}, \ldots A_{t-1}, X_{t-1,A_{t-1}})$. Over a fixed time period (or horizon) of length n, the agent desires to maximize his/her total reward $\sum_{t=1}^n X_{t,A_t}$ by designing an appropriate policy $\pi = \{\pi_t\}_{t=1}^n$ where π_t is $\sigma(\mathcal{H}_{t-1})$-measurable. An equivalent measure is the *regret*

$$R_n(\nu, \pi) := \mathbb{E}\left[n\mu_1 - \sum_{t=1}^n \mu_{A_t}\right],$$

which is to be minimized over π. However, this measure pays no attention to *risk*, which is the central theme of this survey. More generally, the agent would like to minimize the ρ-regret

$$R_n^\rho(\nu, \pi) := \mathbb{E}\left[n \max_{1\leq i\leq K} \rho(\nu_i) - \sum_{t=1}^n \rho(\nu_{A_t})\right],$$

where $\rho(\nu)$ is an appropriate risk measure. Let $\Delta_i = \rho(\nu_{i^*}) - \rho(\nu_i)$ be the gap between the risk of the best arm i^* and that of the i-th arm.

4.1 Confidence Bound-Based Algorithms

We present a straightforward adaptation Risk-LCB of the well-known UCB algorithm [Auer *et al.*, 2002] to handle an objective based on abstract risk measures ρ. The algorithm caters to Lipschitz risk measures (see Definition 5) and arms' distributions that are sub-Gaussian. Here, note that we are *minimizing* the risk measure instead of *maximizing* the reward.

To obtain the confidence widths, we first rewrite the concentration bound in Theorem 1 as follows: For any $\delta \in [\exp(-2m), 1]$, with probability $\geq 1-\delta$,

$$|\rho_m - \rho(X)| \leq L\left[\Big(\frac{256\sigma^2 \mathsf{e}\log(\frac{1}{\delta})}{m}\Big)^{\frac{1}{2}} + \frac{256\sqrt{2}\sigma}{\sqrt{m}}\right],$$

where $\rho_m = \rho(F_m)$ is an m-sample estimate of the risk measure $\rho(X)$. For the classic bandit setup with a expected value objective, the finite sample analysis provided by [Auer *et al.*, 2002] chose to set $\delta = t^{-4}$ for obtaining a sub-linear regret guarantee. This choice also satisfies the constraint on δ specified in the high confidence guarantee on ρ_m above. Under this choice of δ, in any round t of Risk-LCB and for any arm $k \in \{1, \ldots, K\}$, we have

$$\mathbb{P}\left(\rho(i) \in [\rho_{i,T_i(t-1)} \pm w_{i,T_i(t-1)}]\right) \geq 1 - t^{-4},$$

where

$$w_{i,T} := \frac{L\sigma\left[32\sqrt{\mathsf{e}\log t} + 256\sqrt{2}\right]}{T}.$$

In the above, $\rho_{i,T_i(t-1)}$ is the estimate of the risk measure for arm i computed using $\rho_n = \rho(F_n)$ from $T_i(t-1)$ samples.

The Risk-LCB algorithm would play all arms once in the initialization phase, and in rounds $t \geq K+1$, play an arm A_t according to the following rule:

$$A_t = \operatorname*{arg\,min}_{1\leq i\leq K}\left(\mathrm{LCB}_t(i) = \rho_{i,T_i(t-1)} - w_{i,T_i(t-1)}\right).$$

Theorem 2. *Consider a K-armed stochastic bandit problem with a Lipschitz risk measure ρ and the arms' distributions are sub-Gaussian with parameter σ^2. Then, the expected regret R_n^ρ of Risk-LCB satisfies the following bound:*

$$R_n^\rho \leq \sum_{i:\Delta_i>0} \frac{4L^2\sigma^2[32\sqrt{\mathsf{e}\log n}+256\sqrt{2}]^2}{\Delta_i} + 5K\Delta_i.$$

This bound mimics that of risk-neutral UCB except that the Δ_i's depend on ρ.

The regret bound derived in Theorem 2 is not applicable for CPT, as it is not a Lipschitz risk measure. However, one can derive a confidence bound-based algorithm with CPT as the risk measure using the result in Proposition 1 for the case of arms' distributions with bounded support, see Gopalan *et al.* [2017]. The regret bound is of the order $O\big(\sum_{i:\Delta_i>0} \frac{\log n}{\Delta_i^{2/\alpha-1}}\big)$, where α is the Hölder exponent (see **(A1)** above). This bound cannot be improved in terms of dependence on the gaps and horizon n; see Theorem 3 in Gopalan *et al.* [2017] for details. For an extension of confidence bound-based algorithms to handle sub-Gaussian arms' distributions, see Prashanth and Bhat [2020].

4.2 Thompson Sampling

Another popular class of algorithms for regret minimization is *Thompson sampling* [Thompson, 1933; Agrawal and Goyal, 2012; Kaufmann *et al.*, 2012; Russo *et al.*, 2018]. Here, one starts with a prior over the set of bandit environments. In each round, as observations are obtained, the prior is updated to a posterior. The agent then samples an environment from the posterior and chooses the action that is optimal for the sampled environment.

The first attempt at using Thompson sampling for risk-averse bandits was by Zhu and Tan [2020] in which the authors applied Thompson sampling to mean-variance bandits for Gaussian and Bernoulli arms. This work is the Thompson sampling analogue of the UCB-based ones for the mean-variance by Sani *et al.* [2012] and Vakili and Zhao [2015]. For the Gaussian case, Sani *et al.* [2020] considered sampling the precision (inverse variance) of the Gaussian $\tau_{i,t}$ of arm i at time t from a Gamma distribution with parameters $\alpha_{i,t}$ and $\beta_{i,t}$. Subsequently, the mean of the Gaussian $\theta_{i,t}$ is sampled from a Gaussian whose mean is set to be the empirical mean and whose variance is the inverse of the number of times arm i has been played. The Gamma and Gaussian are chosen as they are conjugate to the precision and mean of the Gaussian respectively. This algorithm, termed MVTS (for mean-variance Thompson sampling), has expected regret satisfying

$$\limsup_{n\to\infty} \frac{R_n^\rho}{\log n} \leq \sum_{i=2}^{K} \max\left\{\frac{2}{\Gamma_{1,i}^2}, \frac{1}{h(\frac{\sigma_i^2}{\sigma_1^2})}\right\}(\Delta_i + 2\overline{\Gamma}_i^2),$$

where $\Gamma_{i,j} := \mu_i - \mu_j$ is the gap between the means of arms i and j, $\overline{\Gamma}_i^2 := \max_{j=1,\ldots,K}(\mu_i - \mu_i)^2$, $\Delta_i := \mathrm{MV}_{i^*} - \mathrm{MV}_i$ is the gap between the mean-variance of arm i and the arm with the best MV i^*, and $h(x) := \frac{1}{2}(x - 1 - \log x)$. This bound was shown to be order-optimal in the sense of meeting a problem-dependent information-theoretic lower bound as γ tends to either 0 or diverges to ∞.

Baudry *et al.* [2021] considered an MAB problem in which the quality of an arm is measured by the CVaR of the reward distribution. They proposed B-CVTS and M-CVTS for continuous bounded and multinominal rewards respectively. Leveraging analysis by [Riou and Honda, 2020], the authors bounded the regret performances of both algorithms and show that they are asymptotically optimal. This is desirable as it improves on Zhu and Tan [2020], albeit for different risk measures, since the level α does not have to tend to its extremal values to be asymptotically optimal. Generalizations of and guarantees for Thompson sampling algorithms for other risk measures are discussed in Chang and Tan [2021].

5 Pure Exploration

We now review existing work on risk-aware *pure exploration* in MABs. Risk-aware pure exploration problems typically involve finding the best arm (or the best few arms), where the merit of an arm is defined to accommodate a risk measure or a constraint. Pure exploration problems have been studied in two settings, namely, *fixed budget* and *fixed confidence*.

5.1 Fixed Budget Algorithms

In the fixed budget setting, the objective is to find the arm (or the best few arms) using a fixed number of arm plays, so as to minimize the probability of misidentification. To be more precise, let n be the fixed "budget", i.e., the total number of arm plays allowed. In the notation of the previous section, let $\rho(\nu_k)$ be the risk measure of interest associated with arm k, and let k^* be the optimal arm (often assumed to be unique for simplicity). A given arm selection policy $\pi = \{\pi_t\}_{t=1}^n$ samples the arms as before, and at the end of the budget, identifies the arm k_n^π to be the optimal arm. The merit of the arm selection algorithm is captured by the probability of misidentifying the best arm, which is simply $\mathbb{P}\left(k_n^\pi \neq k^*\right)$.

The recent paper by Zhang and Ong [2021] considers pure exploration in a fixed budget setting, where the risk measure is the VaR or quantile at a particular level. Specifically, given a quantile level α, the objective is to identify a set of m arms with the highest VaR_α values within a fixed budget n. The authors obtain the empirical VaR estimate using the order statistic $X_{(\lfloor n(1-\alpha)\rfloor)}$ as explained in Section 3.3. They then derive exponential concentration bounds for the VaR_α estimate under the assumptions that the underlying r.v.s have an increasing hazard rate, and a continuously differential probability density function. A similar VaR concentration result is derived in Prop. 2 of Kolla *et al.* [2019] under milder assumptions—nevertheless, VaR concentration results necessarily involve constants that depend on the slope of the distribution in the neighbourhood of the α-quantile.

The algorithm proposed in Zhang and Ong [2021] called Q-SAR (Quantile Successive Accept-Reject) adopts the well-known algorithm of Bubeck *et al.* [2013] to the risk-aware setting. Q-SAR first divides the time budget n into $M-1$ phases, where the number of samples drawn for each arm in each phase is chosen exactly as in [Bubeck *et al.*, 2013] for the risk neutral setting. At each phase $p \in \{1, 2, \ldots, M-1\}$, the algorithm maintains two sets of arms, namely the *active set*, which contains all arms that are actively drawn in phase p, and the *accepted set* which contains arms that have been accepted. During each phase, based on the empirical VaR estimates, an arm is removed from the active set, and it is either accepted or rejected. If an arm is accepted, it is added into the accepted set. When the time budget ends, only one arm remains in the active set, which together with the accepted set, forms the recommended set of best arms. In Theorem 4 of Zhang and Ong [2021] the Q-SAR algorithm is shown to have an exponentially decaying probability of error in the budget n, using the VaR concentration bound.

Other papers on risk-aware pure exploration with fixed budget include Kagrecha *et al.* [2019] and Prashanth *et al.* [2020], which consider CVaR as the risk measure. In Prashanth *et al.* [2020], the authors derive concentration bounds for CVaR estimates, considering separately sub-Gaussian, light-tailed (or sub-exponential) and heavy-tailed distributions. For the sub-Gaussian and light-tailed cases, the estimates are constructed as described in Section 3.3. For heavy-tailed random variables, they assume a bounded moment condition, and derive a concentration bound for a truncation-based estimator. For all three distribution classes, the concentration bounds exhibit exponential decay in the number of samples. The authors then consider the best CVaR-arm identification problem under a fixed budget. The algorithm, called CVaR-SR, adopts the well-known successive rejects algorithm from Audibert *et al.* [2010] to best CVaR arm identification. Exponentially decaying bounds on the probability of incorrect arm identification are derived using the CVaR concentration results. We remark that the Wasserstein distance approach outlined in Section 3.2 gives a direct CVaR concentration bound for sub-Gaussian distributions, but does not seem to extend easily for distributions with heavier tails.

In Kagrecha *et al.* [2019], the authors consider a risk-aware best-arm identification problem, where the merit of an arm is defined as a linear combination of the expected reward and the associated CVaR. A notable feature in this paper is the *distribution obliviousness* of the algorithms, i.e., the algorithm (which is again a variant of successive rejects) is not aware of any information on the reward distributions, including tails or moment bounds.

We conclude this subsection by commenting on a common desirable feature in all the algorithms reviewed above. The SR (and SAR) family of algorithms does not require knowledge of the constants in the concentration bounds for their implementation. These constants, which are often distribution dependent, are not known in practice. They appear only in the analysis and the upper bound on the probability of the error. This is in contrast with confidence intervals-based algorithms used for regret minimization, where the distribution dependent constants from the concentration bound do appear in the confidence term, necessitating the knowledge of these constants in the algorithm's implementation.

5.2 Fixed Confidence Algorithms

In fixed confidence pure exploration, the objective is to identify with a high degree of confidence (say with probability at least $1-\delta$) the best arm with the smallest possible number of arm plays. A related (and less stringent) objective is called PAC (Probably Approximately Correct), where the objective is to identify as quickly as possible, the set of arms with reward values that are ϵ-close to the best arm, with a confidence level at least $1-\delta$. To be more precise, a policy in the fixed confidence setting consists of a *stopping time* T (defined w.r.t. the filtration $\sigma(\mathcal{H}_{t-1})$ from the previous section), and an arm selection rule $\{\pi_t\}_{t=1}^T$. The objective is to ensure that the arm identified at the stopping time T is the 'true best arm' k^* with probability at least $1-\delta$. The figure of merit is the expected sample complexity $\mathbb{E}[T]$, which should be as small as possible for a fixed δ.

In Szorenyi *et al.* [2015], the authors consider PAC best arm identification, where the quality of an arm is defined in terms of VaR (or quantile) at a fixed risk level. The key technique therein is to employ a sup-norm concentration of the empirical distribution (an improved version of the DKW inequality [Massart, 1990]) to obtain a suitable VaR concentration result. The authors also provide a lower bound for the special case of the VaR at the level 0.75. A lower bound was then generalized to any level α in David and Shimkin [2016]. Notably, David and Shimkin [2016] also proposes two PAC algorithms (namely Maximal Quantile and Doubled Maximal Quantile) that have sample complexity within logarith-

mic factors of the lower bound.

In David *et al.* [2018], the authors study a PAC problem with risk constraints. Therein, the objective is to find an arm with the highest mean reward (similar to classic best arm identification), but only among those arms that satisfy a risk constraint. The risk measure is defined in terms of the $\mathrm{VaR}_\alpha(\cdot)$ of the arms being no smaller than a given β. The authors propose a confidence bounds-based algorithm and prove an upper bound on its sample complexity for sub-Gaussian arms' distribution. The upper bound has a similar form to the guarantee available for the risk-neutral PAC bandits problem [Even-Dar *et al.*, 2006]. They also show a lower bound on the sample complexity that matches the upper bound within logarithmic factors, using specific problem instances inspired by similar approaches in the risk-neutral setting [Mannor and Tsitsiklis, 2004]. An improvement to the algorithm in David *et al.* [2018] based on the LUCB algorithm [Kalyanakrishnan *et al.*, 2012] was proposed by Hou *et al.* [2022] recently.

6 Future Challenges

The concentration bounds presented herein based on relating the risk measure to Wasserstein distance are likely to be sub-optimal in terms of the constants. One can aim to tighten these bounds, which will have impact and utility beyond the study of MABs. One can explore methods other than the empirical estimators for various risk measures. For example, the potential of using importance sampling for estimating tail-based risk measures such as the CVaR has hitherto not been explored adequately. If improved concentration bounds can be derived, they can be used to obtain improved, problem-dependent bounds in both in regret minimization and best arm identification settings. These will also ascertain the asymptotic optimality of various algorithms by comparing the achievable bounds to the information-theoretic limits. From Section 5, we notice that there is generally much less work on pure exploration with risk, and only VaR and CVaR have been explored. Finally, from a philosophical perspective, it is natural to wonder which risk measure to use when faced with a certain application scenario, such as portfolio optimization or clinical trials. A deeper, quantitative understanding of various risk measures and the ensuing implications on the scenario at hand requires more inter-disciplinary research.

Acknowledgements

VT is supported by a Singapore National Research Foundation (NRF) Fellowship (A-0005077-00-00) and Singapore Ministry of Education AcRF Tier 1 grants (A-0009042-00-00, A-8000189-00-00, and A-8000196-00-00).

References

[Acerbi, 2002] C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

[Agrawal and Goyal, 2012] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

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

[Audibert *et al.*, 2010] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proceedings of the Conference on Learning Theory*, pages 41–53, 2010.

[Auer *et al.*, 2002] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[Baudry *et al.*, 2021] D. Baudry, R. Gautron, E. Kaufmann, and O. Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 716–726. PMLR, Jul 2021.

[Bhat and Prashanth, 2019] S. P. Bhat and L. A. Prashanth. Concentration of risk measures: A Wasserstein distance approach. In *Advances in Neural Information Processing Systems*, pages 11739–11748, 2019.

[Brown, 2007] D. B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

[Bubeck *et al.*, 2013] S. Bubeck, T. Wang, and N. Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265. PMLR, 2013.

[Cassel *et al.*, 2018] A. Cassel, S. Mannor, and A. Zeevi. A general approach to multi-armed bandits under risk criteria. In *Proceedings of the 31st Conference On Learning Theory*, pages 1295–1306, 2018.

[Chang and Tan, 2021] J. Q. L. Chang and V. Y. F. Tan. A unifying theory of thompson sampling for continuous risk-averse bandits. *arXiv preprint arXiv:2108.11345*, 2021.

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

[David *et al.*, 2018] Y. David, B. Szörényi, M. Ghavamzadeh, S. Mannor, and N. Shimkin. PAC bandits with risk constraints. In *ISAIM*, 2018.

[Even-Dar *et al.*, 2006] E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

[Gopalan *et al.*, 2017] A. Gopalan, L. A. Prashanth, M. C. Fu, and S. I. Marcus. Weighted bandits or: How bandits learn distorted values that are not expected. In *AAAI Conference on Artificial Intelligence*, pages 1941–1947, 2017.

[Hou *et al.*, 2022] Y. Hou, V. Y. F. Tan, and Z. Zhong. Almost optimal variance-constrained best arm identification, 2022. arXiv 2201.10142.

[Jie *et al.*, 2018] C. Jie, L. A. Prashanth, M. C. Fu, S. I. Marcus, and C. Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9):2867–2882, 2018.

[Kagrecha *et al.*, 2019] 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*, pages 11269–11278, 2019.

[Kalyanakrishnan *et al.*, 2012] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pages 227–234. PMLR, 2012.

[Kaufmann *et al.*, 2012] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite time analysis. In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 7568, pages 199–213, 2012.

[Knight, 1921] F. H. Knight. *Risk, Uncertainty and Profit*. Houghton Mifflin Company, New York, 1921.

[Kolla *et al.*, 2019] R. K. Kolla, L. A. Prashanth, S. P. Bhat, and K. P. Jagannathan. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47(1):16–20, 2019.

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

[Mannor and Tsitsiklis, 2004] 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.

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

[Massart, 1990] P. Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[Pandey *et al.*, 2021] A. K. Pandey, L. A. Prashanth, and S. P. Bhat. Estimation of spectral risk measures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12166–12173, May 2021.

[Prashanth and Bhat, 2020] L. A. Prashanth and S. P. Bhat. A Wasserstein distance approach for concentration of empirical risk estimates, 2020. arXiv 1902.10709v3.

[Prashanth *et al.*, 2016] L. A. Prashanth, J. Cheng, M. C. Fu, S. I. Marcus, and C. Szepesvári. Cumulative prospect theory meets reinforcement learning: prediction and control. In *International Conference on Machine Learning*, pages 1406–1415. PMLR, 2016.

[Prashanth *et al.*, 2020] L. A. Prashanth, K. Jagannathan, and R. K. Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In *International Conference on Machine Learning*, volume 119, pages 5577–5586. PMLR, 2020.

[Riou and Honda, 2020] C. Riou and J. Honda. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117, pages 777–826. PMLR, Feb 2020.

[Rockafellar and Uryasev, 2000] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

[Russo *et al.*, 2018] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A Tutorial on Thompson Sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

[Sani *et al.*, 2012] A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[Sharpe, 1966] W. F. Sharpe. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966.

[Szorenyi *et al.*, 2015] B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *International Conference on Machine Learning*, pages 1660–1668. PMLR, 2015.

[Thomas and Learned-Miller, 2019] P. Thomas and E. Learned-Miller. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, pages 6225–6233. PMLR, 2019.

[Thompson, 1933] 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.

[Tversky and Kahneman, 1992] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[Vakili and Zhao, 2015] S Vakili and Q. Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335, 2015.

[Wainwright, 2019] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[Wang and Gao, 2010] Y. Wang and F. Gao. Deviation inequalities for an estimator of the conditional value-at-risk. *Operations Research Letters*, 38(3):236–239, 2010.

[Zhang and Ong, 2021] M. Zhang and C. S. Ong. Quantile bandits for best arms identification. In *International Conference on Machine Learning*, pages 12513–12523. PMLR, 2021.

[Zhu and Tan, 2020] Q. Zhu and V. Y. F. Tan. Thompson sampling algorithms for mean-variance bandits. In *International Conference on Machine Learning*, pages 2645–2654. PMLR, 2020.