Sample Complexity of Risk-Averse Bandit-Arm Selection

Jia Yuan Yu* and Evdokia Nikolova†

Abstract

We consider stochastic multiarmed bandit problems where each arm generates i.i.d. rewards according to an unknown distribution. Whereas classical bandit solutions only maximize the expected reward, we consider the problem of minimizing risk using notions such as the value-at-risk, the average value-at-risk, and the mean-variance risk. We present algorithms to minimize the risk over a single and multiple time periods, along with PAC accuracy guarantees given a finite number of reward samples. In the single-period case, we show that finding the arm with least risk requires not many more samples than the arm with highest expected reward. Although minimizing the multi-period value-at-risk is known to be hard, we present an algorithm with comparable sample complexity under additional assumptions.

1 Introduction

Multiarmed bandit problems arise in diverse applications such as tuning parameters, Internet advertisement, auction mechanisms, adaptive routing in networks, project management, and clinical trials. The goal is typically to find a policy, that is, a sequence of decisions, that maximizes the cumulative *expected* reward [Lai and Robbins, 1985]. The essence of the problem, and its principal challenge, lies in the uncertainty of rewards. Due to the risk-averse nature of users in many applications, a solution with guarantees in expectation is often unsatisfactory, *i.e.*, the best solution may be the one with smallest risk as opposed to highest mean.

Risk measures have recently received renewed attention in the financial mathematics and optimization literature [Artzner *et al.*, 1999; Rockafellar, 2007]. However, until now, these risk measures have been rarely applied to multiarmed bandit problems. One reason is that risk analysis often assumes distributional knowledge of the underlying uncertainty, whereas in bandit models, this knowledge is unavailable a priori. The uncertainty must be estimated before making risk-averse decisions. Another reason is that the complexity of estimating risk associated with sequential decision-making increases greatly as the number of decisions increases. This work addresses both of these challenges.

Although the underlying randomness is a priori unknown in many real problems, it is useful to consider notions of risk measures with respect to this unknown randomness. In this work, we consider a data-driven approach to risk aversion, wherein we estimate the true risk measure from a sequence of observations of the randomness. One of our contributions is to quantify the effect of finite samples on the residual risk. We do so for modern and traditional risk measures: the value-at-risk (V@R), the average value-at-risk[1] (AV@R), and the mean-variance risk. This work is the first to apply the AV@R to bandit problems, which is widely used elsewhere because an important class of convex risk measures can be expressed as its integral. Our notion of risk is defined differently from previous work on risk-averse bandits.

When the objective is the expected reward—as in most of the literature on bandit problems, by linearity of expectation, the cumulative expected reward is simply the sum of single-period expected rewards. In contrast, the instantaneous or *single-period* risk and cumulative or *multi-period* risk can vary greatly in complexity. A risk-averse objective is nonlinear and does not typically decompose into a sum over single-period risks. When comparing different arms, it is natural to compare their single-period risks; however, these will not always imply the correct preference with respect to a multi-period risk objective.

Example 1.1 below gives insight into the subtlety of measuring risk over multiple periods. It demonstrates the following counter-intuitive fact: one arm may be better than another in terms of the single period risk , while it may be worse over two or more periods. In addition, even if arm 1 is better than arm 2 in a single period, pulling arm 1 twice may be worse than consecutively pulling arm 1 and arm 2. In contrast to the expected reward criterion, one cannot infer multi-period risk

*IBM Research–Ireland, Damastown, Dublin 15, Ireland, jiayuanyu@ie.ibm.com. This work was supported in part by the EU FP7 project INSIGHT under grant 318225.

†Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843-3112 USA, nikolova@tamu.edu. This work was partly done while the author was visiting IBM Research—Ireland and was supported in part by NSF CCF 1216103.

[1]The average value-at-risk is also called *expected shortfall* and *conditional value-at-risk*.

bounds from single-period risk bounds.

Example 1.1 (Lack of translation symmetry)**.** Consider two arms and two periods. The first arm has rewards X_1, X_2 in periods $1, 2$ respectively and the second arm has rewards Y_1, Y_2. Let X_1, X_2 be independent and identically distributed according to the normal distribution $\mathcal{N}(\mu_1, \sigma_1^2)$, whereas Y_1, Y_2 are i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$. For normal random variables, both the $\mathrm{V@R}_\beta$ and the $\mathrm{AV@R}_\beta$, with probability level $\beta \geqslant 0.5$, reduce to a linear combination of mean and standard deviation: $\rho(X) = -\mathbb{E}[X] + \lambda\sqrt{\mathrm{VAR}[X]}$, where λ is only a function of β (cf. [Rockafellar and Uryasev, 2000, Proof of Proposition 1]). Thus, the single period risk measures are $\rho(X_1) = -\mu_1 + \lambda\sigma_1$ for arm 1 and $\rho(Y_1) = -\mu_2 + \lambda\sigma_2$ for arm 2.

Let $\epsilon > 0$ be fixed. Suppose that $\sigma_1 = 1$, $\sigma_2 = 2$, and $\mu_1 = \lambda$, $\mu_2 = 2\lambda - \epsilon$, then we have $\rho(X_1) = -\mu_1 + \lambda\sigma_1 = 0$ and $\rho(Y_1) = -\mu_2 + \lambda\sigma_2 = \epsilon$. Moreover, by independence, we have $\rho(Y_1 + Y_2) = -2\mu_2 + \lambda\sqrt{2\sigma_2^2} = 2(\sqrt{2} - 2)\lambda + 2\epsilon$, $\rho(X_1 + Y_2) = -\mu_1 - \mu_2 + \lambda\sqrt{\sigma_1^2 + \sigma_2^2} = (\sqrt{5} - 3)\lambda + \epsilon$, $\rho(X_1 + X_2) = -2\mu_1 + \lambda\sqrt{2\sigma_1^2} = (\sqrt{2} - 2)\lambda$. It is easy to verify that, for all $\epsilon > 0$ and $\lambda > 6\epsilon$, we have $\rho(X_1) < \rho(Y_1)$ and $\rho(X_1 + X_2) > \rho(X_1 + Y_2) > \rho(Y_1 + Y_2)$. We conclude that for two periods, we incur less risk (V@R and AV@R) by choosing arm 2 twice; whereas for a single period, it is preferable to choose arm 1.

The following is an engineering example of a bandit problem where risk is important.

Example 1.2 (Communication channel selection)**.** Consider a transmitter with access to a number of communication channels (*e.g.*, different media and spectrum frequencies). Its task is to choose a single channel to transmit one important message so as to minimize the risk corresponding to the probability that the error rate exceeds a threshold λ—instead of minimizing the expected error rate. The channel error rates are random with unknown distribution, but can be sampled. The question is: How many samples are required to find a channel satisfying some prescribed confidence guarantees on its risk?

We proceed as follows. In Sections 2 and 3, we present our bandit model and discuss it with respect to related literature. In Section 4, we present PAC bounds on the single-period risk of a greedy arm-selection policy with respect to an appropriate risk estimate. Section 5 presents our main result: a PAC bound for multi-period risk using a new algorithm and under an additional assumption. Section 6 illustrates empirically the distinction between risk-averse and expected-reward bandit problems. We discuss open problems in Section 7.

2 Problem formulation

Let $\{1, \dots, n\}$ denote a set of arms—or possible choices of actions. Let $\{X_t^i : i = 1, \dots, n, t = 1, 2, \dots\}$ denote the real-valued rewards for pulling the arms $i = 1, \dots, n$ at time instants $t = 1, 2, \dots$ Let $d_1, \dots, d_n$ denote time-invariant probability density functions. We assume that for every fixed arm i, the rewards $\{X_1^i, X_2^i, \dots\}$ are independent and identically distributed according to d_i. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space generated by the arm rewards. In contrast to most of the literature on risk measures, we assume that $\mathbb{P}$—likewise the densities $d_1, \dots, d_n$—are fixed but *unknown*.

Let i_t denote the action chosen by the decision-maker at time t. Let μ_i denote the expected reward of arm i—with density d_i. In the traditional bandit problem formulation, we are interested in algorithms for choosing the sequence of arms $\{i_t\}$ with expected regret guarantees of the form

$$\left| \max_{i=1,\dots,n} \mu_i - \mathbb{E}\, \frac{1}{T} \sum_{t=1}^{T} X_t^{i_t} \right| \in o\left(\frac{\log T}{T}\right).$$

Let $\mathcal{Y}$ denote the set of bounded random variables, and let $\rho : \mathcal{Y} \to \mathbb{R}$ denote a risk measure. We discuss specific types of risk measures in the coming sections. We are interested in a risk-averse version of bandit problems, where one objective is to give a PAC bound on the *single-period risk error* $\left|\min_{j=1,\dots,n} \rho(X_1^j) - \rho(X_1^{i^*})\right|$ of the chosen arm i^*. Another objective is to give a PAC bound on the *multi-period risk error* of a sequence of actions $i_1, \dots, i_\tau$:

$$\left| \min_{(a_1,\dots,a_\tau)\in[n]^\tau} \rho\left(\sum_{t=1}^{\tau} X_t^{a_t}\right) - \rho\left(\sum_{t=1}^{\tau} X_t^{i_t}\right) \right|. \quad (1)$$

In this paper, our notion of multi-period risk assumes that the arm choices for time instants $1, \dots, \tau$ are made at time 1, while ignoring outcomes after time 1. An alternative would be to consider risk associated with policies that act at time t according to all observed outcomes up to time $t-1$. Under some assumptions—particularly, when the decision horizon τ is short compared to the number of past observations, these risk measures are close.

Remark 1 (Notation). Since $\{X_1^i, X_2^i, \dots\}$ are identically distributed, we have $\rho(X_1^i) = \rho(X_2^i) = \dots$, and we write $\rho(i)$ to denote $\rho(X_1^i)$.

For every fixed arm i, let $\mathcal{X}^i$ denote a set of observations or samples of the rewards of arm i. Throughout the paper, we assume that these sets contain the same number N of samples, *i.e.*, $\left|\mathcal{X}^1\right| = \dots = |\mathcal{X}^n| = N$. The number N is called the *sample complexity*; it represents the number of samples that is sufficient to give a certain guarantee.

3 Related work

The notion of risk has been widely studied in finance, engineering, and optimization, yet there is no single agreed-upon definition for it. It is generally meant to express and quantify one's preferences over a set of random outcomes. Two main approaches to modeling risk have been via utility functions [Neumann and Morgenstern, 1944] and via the mean-variance framework [Markowitz, 1952]. More recently, due to a variety of paradoxes and pitfalls of the traditional approaches, an axiomatic approach to risk has been proposed for applications in finance [Artzner *et al.*, 1999]. This approach to risk-aversion spans both static and sequential decision problems, such as [Artzner *et al.*, 1999] in financial hedging; [Le Tallec, 2007; Osogami, 2011] in Markov decision problems; [Shapiro and Ahmed, 2004] in stochastic and robust optimization. Both the traditional and modern approaches have been comprehensively described in multiple surveys (cf. [Schied, 2006; Rockafellar, 2007]).

Multiarmed bandit problems have been studied in a variety of settings, including the Markovian (rested and restless), stochastic and adversarial settings. For surveys on bandit problems, we refer the reader to [Gittins *et al.*, 2011; Cesa-Bianchi and Lugosi, 2003]. Two types of results are found in the literature: results on the average regret (i.e., in a regret-minimization setting [Lai and Robbins, 1985]) and results on the sample complexity (i.e., in a pure-exploration setting [Even-Dar *et al.*, 2002; Bubeck *et al.*, 2011]). Our work is of the second type. It is related to work on sample complexity of bandit arm-selection [Even-Dar *et al.*, 2002; Kalyana *et al.*, 2012], which is also known as pure exploration or best-arm identification [Audibert *et al.*, 2010; Gabillon *et al.*, 2012].

As far as we know, this is the first work to consider the sample complexity of bandit arm-selection in a risk-averse setting. Risk aversion in bandit problems has previously been studied in two settings, Markovian [Gittins *et al.*, 2011] and stochastic [Lai and Robbins, 1985]. In the Markovian setting, [Denardo *et al.*, 2007; Chancelier *et al.*, 2009] consider a one-armed bandit problem in the setting of Gittins indices and model risk with concave utility functions. In the stochastic setting, the notion of risk has been limited to empirical variance [Audibert *et al.*, 2009; Sani *et al.*, 2012]. Besides this limitation, our notion of risk is also very different from this previous work in its definition. In [Audibert *et al.*, 2009; Sani *et al.*, 2012], the risk measure assigns real values to the decision-maker's *policies* (*i.e.*, confidence-bound algorithms) and guarantees are given for the *regret* in retrospect. Our risk measure assigns a real value to random variables, *i.e.*, *rewards* of individual arms or deterministic sequences of arm choices, and guarantees are given on the sample complexity of various estimators. This is more in line with the risk notions of the finance and optimization literature. Our results on the sample complexity of estimating the V@R and the AV@R are new. Our results on the mean-variance risk complement those of [Sani *et al.*, 2012].

Our notion of data-driven risk, estimated using random samples, is similar to [Jones and Zitikis, 2003; Brown, 2007; Kim and Hardy, 2009]: [Jones and Zitikis, 2003] presents an empirical study of bootstrapped risk estimators; we present an alternative AV@R estimator to that of [Brown, 2007] that has comparable sample-complexity guarantees, but is more efficient by virtue of not solving a minimization problem; [Kim and Hardy, 2009] estimate tail-deviation risk measures using L-statistics and show asymptotic properties of the estimators. In this paper, we consider different risk measures and present non-asymptotic results. Our sample complexity results on estimating risk measures is also reminiscent of black-box models [Shmoys and Swamy, 2006; Nemirovski and Shapiro, 2006]. However, the notion of risk-aversion in [Shmoys and Swamy, 2006] is limited to V@R and the setting is that of the two-stage recourse model. For [Nemirovski and Shapiro, 2006], the setting is chance-constrained optimization and the sample complexity is analyzed for given confidence and reliability parameters.

Even when the underlying probability distribution is known, some risk measures are hard to compute and approximations are used. For instance, methods such as parametric estimation, Monte Carlo simulation, etc. have been used extensively to approximate the risk measures (cf. [Kreinin *et al.*, 1998]). [Vanduffel *et al.*, 2002] presents techniques for approximating risk measures of a sum of random variables with known lognormal distributions. In our setting, the multi-period risk is defined for random variables with arbitrary and unknown distributions.

In this paper, we first want to raise awareness of the need to incorporate risk in machine learning problems—the experiments of Section 6 illustrate this. Secondly, we investigate how risk affects the underlying theory compared to the classical expected-reward setting. We show where existing techniques can give sample complexity guarantees in the new risk-averse formulations and where new techniques are needed. In particular, we present the gap in complexity between the single-period and the multi-period risk-averse bandit problems. The latter is hard even with an additional assumption of independence between arms. Our main contribution is to present an arm-selection algorithm (the CuRisk algorithm) with a PAC guarantee on its multi-period risk.

4 Single-period risk

In this section, we present results on the sample complexity of estimating single-period risk, and derive PAC bounds for the single-period best-arm identification problem. The results for the mean-variance risk can be deduced from well-known results in the literature. We present them separately because they do not require an assumption of independence between arms (cf. Assumption 5.1). We consider the multi-period case in the next section.

4.1 Value-at-risk

Let λ be given and fixed. In this section, we consider the value-at-risk, for every arm i:

$$\rho_\lambda^V(i) = \mathrm{V@R}_\lambda(i) = -q_i(\lambda),$$

where q_i is the right-continuous quantile function[2] of X_1^i.

Suppose that up to time T, each arm is sampled N times. Let $X_1^i, \ldots, X_N^i$ denote the sequence of rewards generated by arm i. Let $X_{(1)}^i \leqslant \ldots \leqslant X_{(N)}^i$ denote a reordering of the random variables $\{X_1^i, \ldots, X_N^i\}$, where $X_{(k)}^i$ is the k-th order statistic of the sequence $\{X_1^i, \ldots, X_N^i\}$. We consider the following V@R estimators for all i:

$$\widehat{\rho}_\lambda^V(i) \triangleq -\widehat{X}_\lambda^i,$$

where each $\widehat{X}_\lambda^i$ is the following λ-quantile estimator[3]:

$$\widehat{X}_\lambda^i \triangleq X_{(\lceil \lambda N \rceil)}^i. \tag{2}$$

We define our V@R estimator as $\widehat{\rho}_\lambda^V(i) = -\widehat{X}_\lambda^i$.

Assumption 4.1 (Differentiable reward density)**.** For each arm i, the reward probability density functions d_i are continuously differentiable.

[2]Formally, $q_i(\lambda) = \inf\{x \in \mathbb{R} : F_i(x) > \lambda\}$, where F_i is the distribution function of X_1^i.

[3] With slight modifications, we can derive similar results with other quantile estimators, such as the Wilks estimator.

Next, we bound the single-period risk error of the arm-choice $i^* \in \arg\min_{i=1,\ldots,n} \rho^V_\lambda(i)$. Recall that although $\rho^V_\lambda(i)$ is a scalar for every fixed i, the arm-choice i^* is a random variable and $\rho^V_\lambda(i^*)$ is a function of i^*. All proofs appear in the appendix of [Yu and Nikolova, 2013].

Theorem 4.1 (V@R PAC bound)**.** *Suppose that Assumption 4.1 holds. Suppose that there exist D and D' such that $d_i(z) \leqslant D$ and $d'_i(z) \leqslant D'$ for all $z \in \mathbb{R}$ and all i, and that*

$$N \geqslant \max\left\{ \frac{2n\lambda(1-\lambda)}{\delta\varepsilon^2\, D^2}, \sqrt{\frac{16C_2 n}{\delta\varepsilon^2}}, \frac{2\lambda(1-\lambda)D'}{\varepsilon\, D^3}, \sqrt{\frac{4C_1}{\varepsilon}} \right\}.$$

If $i^* \in \arg\min_{i=1,\ldots,n} \widehat{\rho}^V_\lambda(i)$, then arm i^* is (ε,δ)-optimal with respect to the V@R *risk measure,* i.e., $\left|\min_{j=1,\ldots,n} \rho^V_\lambda(j) - \rho^V_\lambda(i^*)\right| \leqslant \varepsilon$ *w.p.* $1-\delta$.

Remark 2. In Theorem 4.1, the sample complexity T is of the order of $\Omega(n\varepsilon^{-2}\delta^{-1})$. By comparison, the sample complexity for the bandit problem with expected rewards is of the order of $\Omega(\varepsilon^{-2}\log(\delta^{-1}))$ [Even-Dar *et al.*, 2002]. We can easily improve the suboptimal dependence in the number of arms n by using the arm-elimination method of [Even-Dar *et al.*, 2002]. Our V@R estimator is very distinct from the empirical mean or distribution estimates in [Even-Dar *et al.*, 2002] and elsewhere in the literature, which allow the use of particular concentration inequalities to obtain a logarithmic dependence in δ^{-1}.

Remark 3 (Lower bound)*.* If X is a Bernoulli random variable with mean p, then the λ-quantile of X is $q_X(\lambda) = 1_{[\lambda \geqslant p]}$, so that estimating the quantile is at least as hard as estimating the mean p. Hence, we can derive a lower bound of $\Omega(n\varepsilon^{-2}\log\delta^{-1})$ on the sample complexity by using the approach of [Mannor and Tsitsiklis, 2004].

4.2 Average value-at-risk

Modern approaches to risk measures [Artzner *et al.*, 1999] advocate the use of convex risk measures, which capture the fact that diversification helps reduce risk. In this section, we consider only one instance of convex risk measures: the average value-at-risk. Nonetheless, it can be shown that an important subset of convex risk measures (*i.e.*, those continuous from above, law invariant, and coherent) can be expressed as an integral of the AV@R (cf. [Schied, 2006]). Guarantees can be obtained for those risk measures by using the approach of this section.

The AV@R has the following two equivalent definitions—first, as an integral of V@R:

$$\rho^A_\lambda(X) = \mathrm{AV@R}_\lambda(X) = \frac{1}{\lambda}\int_0^\lambda \mathrm{V@R}_\phi(X)\mathrm{d}\phi,$$

and second, as a maximum over a set of distributions: $\rho^A_\lambda(X) = \max_{Q \in \mathcal{Q}_\lambda(\mathbb{P})} -\mathbb{E}_Q X$, where $\mathcal{Q}_\lambda(\mathbb{P})$ is the set of probability measures $\{Q : \frac{\mathrm{d}Q}{\mathrm{d}\mathbb{P}} \leqslant 1/\lambda\}$. Depending on the choice of definition, we can estimate the AV@R either via quantile estimation or density estimation. In this section, we adopt the first definition and introduce the following estimator using a Riemann sum of quantile estimator of (2):

$$\widehat{Y}^i_\lambda \triangleq \frac{1}{\lambda}\left(\sum_{j=0}^{\lfloor \lambda N \rfloor - 1} \frac{1}{N} X^i_{(j+1)} + \left(\lambda - \frac{\lfloor \lambda N \rfloor}{N}\right) X^i_{(\lceil \lambda N \rceil)} \right).$$

We define our AV@R estimator as $\widehat{\rho}^A_\lambda(i) = -\widehat{Y}^i_\lambda$. Our estimator is distinct from and computationally more efficient than the one introduced in [Brown, 2007].

The following result bounds the single-period risk of a simple arm-selection policy.

Theorem 4.2 (AV@R PAC bound)**.** *Suppose that the assumptions of Theorem 4.1 hold. Suppose that the rewards are bounded such that $\left|X^i_t\right| \leqslant M$ almost surely, for every arm i and time t. Suppose that*

$$N \geqslant \max\left\{ \frac{32\lambda' M^2}{\varepsilon^2\lambda^2}\log(4n/\delta), \frac{(1/6)D'/D^3 + 2C_1\lambda'}{\varepsilon\lambda}, 2 \right\},$$

where λ' denotes the smallest real number greater than λ such that $\lambda' N$ is an integer. If $i^* \in \arg\min_{i=1,\ldots,n} \widehat{\rho}^A_\lambda(i)$, then arm i^* is (ε,δ)-optimal with respect to the AV@R *risk measure,* i.e., $\left|\min_{j=1,\ldots,n} \rho^A_\lambda(j) - \rho^A_\lambda(i^*)\right| \leqslant \varepsilon$ *w.p.* $1-\delta$.

4.3 Mean-variance risk

In this section, we consider the mean-variance risk measure

$$\rho^M_\lambda(i) = -\mu(i) + \lambda\sigma^2(i),$$

where $\mu(i)$ and $\sigma^2(i)$ denote the mean and variance of arm i. The mean-variance risk measure has been used in risk-averse problem formulations in a variety of applications in finance and reinforcement learning [Markowitz, 1952; Mannor and Tsitsiklis, 2011; Sani *et al.*, 2012].

Let λ be given and fixed. We assume that we are given N samples $X^i_1, \ldots, X^i_N$ for every arm i. We look at the mean-variance risk measure. We employ the risk estimate

$$\widehat{\rho}^M_\lambda(i) = -\widehat{\mu}(i) + \lambda\widehat{\sigma}^2(i),$$

where $\widehat{\mu}(i)$ and $\widehat{\sigma}^2(i)$ denote the sample-mean and the unbiased variance estimator $\widehat{\sigma}^2(i) = \frac{1}{N-1}\sum_{k=1}^N (X^i_k - \widehat{\mu}(i))^2$.

The following theorem gives a single-period risk error bound for the arm-choice $i^* \in \arg\min_{i=1,\ldots,n} \widehat{\rho}^M_\lambda(i)$.

Theorem 4.3 (MV PAC bound)**.** *Suppose that there exist A, B such that $\mathbb{P}(X^i_t \in [A,B]) = 1$ for all i. Suppose that N is at least*

$$\max\left\{ \frac{(B-A)^2}{2\varepsilon^2}\log(8n/\delta), \frac{(B-A)^4\lambda^2}{\varepsilon^2}\log(8n/\delta) + 1 \right\}.$$

If $i^* \in \arg\min_{i=1,\ldots,n} \widehat{\rho}^M_\lambda(i)$, then arm i^* is (ε,δ)-optimal with respect to the mean-variance risk, i.e., $\left|\min_{j=1,\ldots,n} \rho^M_\lambda(j) - \rho^M_\lambda(i^*)\right| \leqslant \varepsilon$ *w.p.* $1-\delta$.

5 From single- to multi-period risk

Obtaining multi-period risk bound estimates from single-period risk is not obvious (cf. Example 1.1). We present in this section multi-period risk bounds for the V@R. The same method can be extended to other risk measures. For instance, by employing the averaging and Riemann approximation method of Section 4.2, we can obtain multi-period risk bounds for the AV@R.

One naive approach to compare the risk of two consecutive actions is to estimate $\rho(X_1 + \ldots + X_\tau)$ for all sequences of arm-choices, but this method requires a number of samples exponential in the time-horizon. We can drastically cut the number of samples when the arms' rewards are mutually independent. In that case, we can estimate the density of $X_1 + \ldots + X_\tau$ by constructing histogram density estimates for each X_i individually, performing the convolution of these histograms (in the transform domain) to obtain a density estimate for $X_1 + \ldots + X_\tau$, and then computing quantile estimates from the latter. Hence, in this section, we make the additional assumption that the rewards are independent across different arms. The exposition is also simplified by assuming that the rewards take a finite number of values instead of Assumption 4.1.

Assumption 5.1 (Arm-wise independence)**.** For every fixed time t, the arms rewards $\{X_t^1, \ldots, X_t^n\}$ are independent.

Assumption 5.2 (Discrete-valued rewards)**.** Let K be a fixed integer. For every arm i, the rewards X_t^i take values in a finite set $\{v_k \in \mathbb{R} : k = 1, \ldots, K\}$ and there exists a constant $\gamma \triangleq \gamma_K > 0$ such that the probability mass function satisfies $d_i(k) \geqslant \gamma$ for all $k = 1, \ldots, K$.

The results of this section extend easily to the case where the rewards take values in interval $[A, B]$; in this case, we partition the subset $[A, B]$ into K intervals $u_1, \ldots, u_K$ such that each interval has length $|B - A| / K$, and take the midpoint of each interval.

We begin by considering the special case of the mean-variance risk measure ρ_λ^M. We can easily derive a multi-period PAC risk bound, provided that the rewards are independent between different arms.

Corollary 5.1 (Multi-period MV PAC bound)**.** *Suppose that the assumptions of Theorem 4.3 hold. Suppose further that Assumption 5.1 holds. If the arm* $i^* \in \arg\min_{i=1,\ldots,n} \rho_\lambda^M(i)$ *is chosen repeatedly at time periods* $1, \ldots, \tau$*, then, with probability* $1 - \delta$*, we have*

$$\left| \min_{(a_1,\ldots,a_\tau)\in[n]^\tau} \rho_\lambda^M \left(\sum_{t=1}^{\tau} X_t^{a_t} \right) - \rho_\lambda^M \left(\sum_{t=1}^{\tau} X_t^{i^*} \right) \right| \leqslant \tau\varepsilon.$$

However, for more general risk measures, more complex techniques are needed to bound the multi-period risk. For the case of the V@R risk measure, we present the CuRisk Algorithm of Figure 1. This algorithm first estimates the probability density of each arm's reward, then solves an allocation problem with respect to the risk measure of interest. The algorithm can be adapted to other risk measures by replacing the ALC-VAR step with corresponding allocation problems. For instance, we can modify the CuRisk Algorithm to tackle the case of the AV@R risk measure using the method of Section 4.2.

The CuRisk Algorithm works as follows. For every arm i, $\widehat{d}_i$ denotes the empirical probability mass function or histogram. The function $\widehat{D}_j$ in Eq. (3) denotes the probability-generating function of X_1^j, which is the z-transform of the probability-mass function $\widehat{d}_j$. The decision variables of the optimization problem ALC-VAR are integers $m_1, \ldots, m_n$ and a real number x. The damping parameter r of the CuRisk Algorithm affords a trade-off between round-off errors in the constraints of ALC-VAR and the generality of the following multi-period risk bound.

Input: A set of arms $\{1, \ldots, n\}$, integers $N > 0$ and $K > 0$, scalar $r \in (0, 1)$, and sets of reward samples $\mathcal{X}^1, \ldots, \mathcal{X}^n$.
Output: Arm choices $i_1, \ldots, i_\tau$.
for all $i = 1, \ldots, n$ **do**
  Compute empirical histogram estimates $\widehat{d}_i$:

$$\widehat{d}_i(k) = \frac{1}{|\mathcal{X}^i|} \sum_{X \in \mathcal{X}^i} 1_{[X \in u_k]}, \quad \text{for all } k = 1, \ldots, K.$$

end for
Solve the allocation problem (ALC-VAR):

$$\begin{aligned} \max_{m_1,\ldots,m_n \in \mathbb{N}_+} \sup_{x \in \mathbb{R}} \quad & x \\ \text{s.t.} \quad & \sum_{k=1}^{\lfloor xK \rfloor} \frac{1}{2kr^k} \sum_{j=1}^{2k} (-1)^j \Re \left[\prod_{\ell=1}^{n} \widehat{D}_\ell^{m_\ell}(re^{\iota j\pi/k}) \right] \leqslant \lambda, \\ & m_1 + \ldots + m_n = \tau, \end{aligned}$$

where $\iota = \sqrt{-1}$, $\Re$ denotes the real-part operator, and $\widehat{D}_\ell$ is the z-transform:

$$\widehat{D}_\ell(z) \triangleq \sum_{k=1}^{K} \widehat{d}_\ell(k) z^k, \quad \text{for all } z \in \mathbb{C}. \tag{3}$$

for $t = 1, \ldots, \tau$ **do**
  Output each arm j exactly m_j^* times, where $(m_1^*, \ldots, m_n^*)$ is a solution to ALC-VAR.
end for

Figure 1: CuRisk Algorithm for V@R

One of the features of the CuRisk Algorithm is that it does not estimate the risk for each possible sequence of actions $\vec{a}$, since this would require a number of times exponential in the length of $\vec{a}$. Instead, the CuRisk Algorithm solves an optimization problem ALC-VAR. Another important feature of the CuRisk Algorithm is that the computational complexity does not increase with the time horizon τ, which is guaranteed in the following theorem. This feature comes from the use of the product of probability-generating functions to approximate a convolution of τ probability mass functions.

Theorem 5.2 (Multi-period V@R PAC bound)**.** *Suppose that Assumptions 5.1 and 5.2 hold, and that*

$$\begin{aligned} K &\geqslant \frac{(\lambda + \gamma)(1 - r^2)^2 + 2}{\varepsilon\gamma(1 - r^2)^2}, \\ N &\geqslant \frac{32\tau^2}{(K\gamma\varepsilon - \lambda - \gamma)^2} \log \left(\frac{4 \cdot 2^K \tau n^\tau}{\delta} \right). \end{aligned}$$

Suppose that the arm choices $i_1, \ldots, i_\tau$ *follow the CuRisk Al-*

gorithm. Then, with probability $1-\delta$*, we have*

$$\left| \min_{(a_1,\ldots,a_\tau)\in[n]^\tau} \rho^V_\lambda \left(\sum_{t=1}^{\tau} X_t^{a_t}\right) - \rho^V_\lambda \left(\sum_{t=1}^{\tau} X_t^{i_t}\right) \right| \leqslant 2\varepsilon.$$

Although we have bounded the sample-complexity of the multi-period problem, the computational complexity remains high due to the non-linear constraint and integer-valued decision variables of ALC-VAR. A possible relaxation is to first solve the real-valued problem and follow-up by rounding.

6 Empirical results

In this section, we first show how different risk measures induce different optimal decisions in the single-period case. Then, we illustrate the complexity of the multi-period case. Details of these examples appear in the appendix of [Yu and Nikolova, 2013].

For the single-period case, we consider a set of five arms with different reward distributions, but the same mean. The reward distributions are: uniform, normal, exponential, Pareto, and Beta. Figure 2 illustrates the convergence of the single-period AV@R estimates for an arm with Pareto reward distributions. Figure 3 shows how different arms are optimal for a single period with respect to different risk measures.

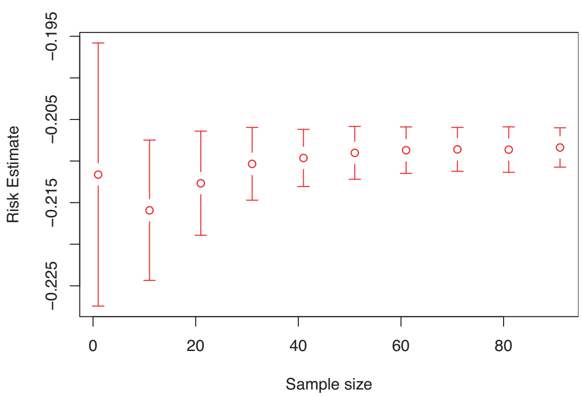

Figure 2: Convergence of the AV@R estimator $\widehat{\rho}^A_\lambda$ for Pareto rewards with error bars at one standard-deviation.

Next, we consider a three-period case with three arms. The first arm generates i.i.d. rewards with values -10 and -20 with probabilities 0.9 and 0.1; the second arm values -5 and -15 with equal probability 0.5; the third arm generates a deterministic reward -14. Figure 4 shows that choosing each of the three arms once minimizes the V@R for a parameter $\lambda \in (0.05, 0.1)$. This shows that non-intuitive solutions can arise from even the simplest of multi-period risk-minimization problems.

7 Discussion

To summarize, for single-period risk, $\Omega(n^2\delta^{-1}\varepsilon^{-2})$ samples are sufficient for a simple policy to find an (ε,δ)-optimal arm with respect to the V@R and AV@R. With respect to the mean-variance risk, $\Omega(n\log(n\delta^{-1})\varepsilon^{-2})$ samples are sufficient. For the multi-period risk over τ periods, $\Omega(n\tau^2\log(\tau n^\tau\delta^{-1})\varepsilon^{-2})$ samples are sufficient for the

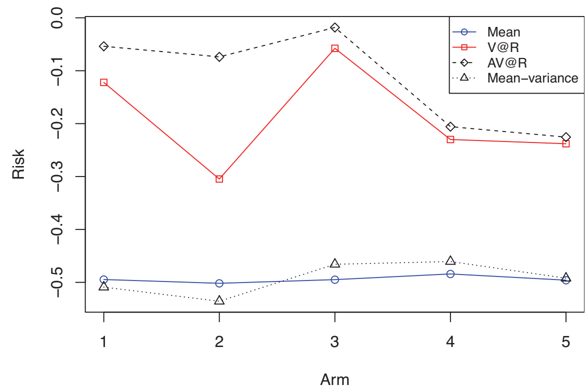

Figure 3: Single-period risk of different arms for different risk measures with $\lambda = 0.1$.

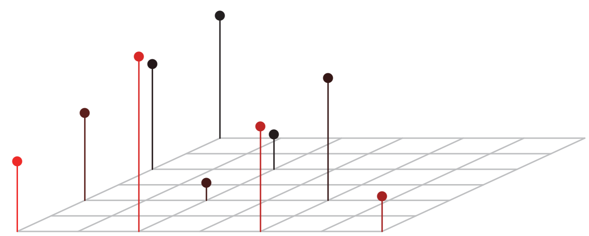

Figure 4: Each bar corresponds to one of ten possible selections among three arms over three periods; the height of each bar corresponds to the associated V@R risk with $\lambda = 0.07$. Each point (a, b) in the (x, y)-plane corresponds to selecting the first, second, and third arm a, b , and $3-a-b$ times respectively. The least V@R selection of arms corresponds to selecting each arm once (*i.e.*, the point $(1, 1)$ in the plane).

CuRisk Algorithm to find an (ε,δ)-optimal sequence of τ arm choices. Using sampling methods that quickly eliminate arms with high risk and low variance [Even-Dar *et al.*, 2002], it is possible to reduce the dependence in the number n of arms.

A number of questions remain open. Is it possible to devise an algorithm with similar PAC risk bounds that do not depend on the problem parameters D and D' of Theorem 4.1? This could, for instance, be achieved by estimating these parameters in parallel. What are lower bounds on the sample complexity for various risk measure estimates? Can we extend our results for V@R and AV@R to the setting of [Sani *et al.*, 2012], where risk measures are associated with policies instead of fixed sequences of actions? A natural policy is one to act greedily according to risk estimates that are updated after each reward observation.

In our analysis, we consider separately the risk due to exploration (or estimation) and the risk due to randomness. It would be interesting to introduce a new notion of risk that combines the two risks. For example, we can define a new risk measure as a weighted sum of two components: the risk with respect to an estimated distribution and the risk of the estimation error.

References

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

[Audibert *et al.*, 2009] J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902, 2009.

[Audibert *et al.*, 2010] J.Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proceedings of COLT*, 2010.

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

[Bubeck *et al.*, 2011] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuously-armed bandits. *Theor. Comput. Sci.*, 412:1832–1852, 2011.

[Cesa-Bianchi and Lugosi, 2003] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Mach. Learn.*, 51(3):239–261, 2003.

[Chancelier *et al.*, 2009] J.-P. Chancelier, M. De Lara, and A. de Palma. Risk aversion in expected intertemporal discounted utilities bandit problems. *Theory and Decision*, 67(4):433–440, 2009.

[Denardo *et al.*, 2007] E. V. Denardo, H. Park, and U. G. Rothblum. Risk-sensitive and risk-neutral multiarmed bandits. *Math. Oper. Res.*, 32(2):374–394, 2007.

[Even-Dar *et al.*, 2002] E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *COLT*, 2002.

[Gabillon *et al.*, 2012] V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck. Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS*, 2012.

[Gittins *et al.*, 2011] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed Bandit Allocation Indices*. Wiley, 2011.

[Jones and Zitikis, 2003] B. L. Jones and R. Zitikis. Empirical estimation of risk measures and related quantities. *N. Am. Actuar. J.*, 7:44–54, 2003.

[Kalyana *et al.*, 2012] S. Kalyana, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *ICML*, 2012.

[Kim and Hardy, 2009] J. H. T. Kim and M. R. Hardy. Estimating the variance of bootstrapped risk measures. *ASTIN Bulletin*, 39(1):199–223, 2009.

[Kreinin *et al.*, 1998] A. Kreinin, L. Merkoulovitch, D. Rosen, and M. Zerbs. Measuring portfolio risk using quasi Monte Carlo methods. *Algo. Res. Quarterly*, 1(1), 1998.

[Lai and Robbins, 1985] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[Le Tallec, 2007] Y. Le Tallec. *Robust, Risk-Sensitive, and Data-driven Control of Markov Decision Processes*. PhD thesis, MIT, 2007.

[Mannor and Tsitsiklis, 2004] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *JMLR*, 5:623–648, 2004.

[Mannor and Tsitsiklis, 2011] S. Mannor and J. N. Tsitsiklis. Mean-variance optimization in markov decision processes. In *ICML*, 2011.

[Markowitz, 1952] H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–98, 1952.

[Nemirovski and Shapiro, 2006] A. Nemirovski and A. Shapiro. *Probabilistic and Randomized Methods for Design under Uncertainty*, chapter Scenario Approximations of Chance Constraints. Springer, 2006.

[Neumann and Morgenstern, 1944] J. Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton U. Press, 1944.

[Osogami, 2011] T. Osogami. Iterated risk measures for risk-sensitive markov decision processes with discounted cost. In *Uncertainty in Artificial Intelligence*, 2011.

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

[Rockafellar, 2007] R. T. Rockafellar. Coherent approaches to risk in optimization under uncertainty. *Tutorials in Operations Research*, pages 38–61, 2007.

[Sani *et al.*, 2012] A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *ICML Workshop*, 2012.

[Schied, 2006] A. Schied. Risk measures and robust optimization problems. *Stoch. Models*, 22(4):753–831, 2006.

[Shapiro and Ahmed, 2004] A. Shapiro and S. Ahmed. On a class of minimax stochastic programs. *SIAM J. Optim.*, 14(4):1237–1249, 2004.

[Shmoys and Swamy, 2006] D. B. Shmoys and C. Swamy. An approximation scheme for stochastic linear programming and its application to stochastic integer programs. *Journal of the ACM*, 2006.

[Vanduffel *et al.*, 2002] S. Vanduffel, T. Hoedemakers, and J. Dhaene. Comparing approximations for risk measures of sums of non-independent lognormal random variables. *N. Amer. Actuarial J.*, 9(4):71–82, 2002.

[Yu and Nikolova, 2013] J. Y. Yu and E. Nikolova. Sample complexity of risk-averse bandit-arm selection. Technical report, IBM, 2013.