

# Generalized Risk-Aversion in Stochastic Multi-Armed Bandits

Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee

Institute of Science and Technology Austria  
{azimin,ribsen,krishnendu.chatterjee}@ist.ac.at

**Abstract.** We consider the problem of minimizing the regret in stochastic multi-armed bandit, when the measure of goodness of an arm is not the mean return, but some general function of the mean and the variance. We characterize the conditions under which learning is possible and present examples for which no natural algorithm can achieve sublinear regret.

## 1 Introduction

The stochastic multi-armed bandit problem is a well-studied framework to model sequential decision-making problems. It has a wide range of theoretical as well as practical applications such as clinical trials, web advertisement placement, packet routing, to name a few. In the usual formulation, an agent (a learner, or an algorithm) has to choose from one of several unknown distributions (which are called arms), receive a sample (a loss) from the arm chosen, and repeat this process for some prescribed amount of time. The goal of the learner is expected regret minimization, i.e., minimization of the expectation of the difference between its own cumulative loss and the cumulative loss of the best arm, where the best arm is the one with the smallest mean. However, for some applications the expected criterion might not be the most desirable. For example, in clinical trials one might not be interested in the most effective treatment on average, but in the one that is more robust and still has a good effect on average. In terms of multi-armed bandits, in this case the best arm is defined not by the mean, but by some risk measure, which is a function of the distribution itself. This leads to the idea of the risk-averse bandit problem.

Risk-aversion has been extensively studied in other fields. Starting from the economic theory ([12], [20]) and ending up with the neighbouring field of reinforcement learning ([6], [19], [18], [13]). In the field of online learning, risk-aversion was studied in the experts setting by [7]. They obtained several negative and positive results for when Sharpe-ratio ([17]) and mean-variance ([12]) was used as risk measures. [21] studied the problem of pure variance minimization. Other risk measures were studied in [16] and [10]. The former proposes to use the mean-variance criterion as a measure of risk and aims at minimizing the notion of the regret that takes into account the variability of the algorithm. The

latter considers log-exponential risk measure, which belongs to the class of so-called coherent risk measures ([14]) and minimizes the regret defined using this measure.

There is no universally agreed notion of what a good measure of risk is, and the appropriate notion can vary from one problem to another. All previous works focused on some particular risk measures, which has immediately limited the applicability of the results and raised a lot of questions on the quality of the particular risk measure. In this work, we consider a different approach: instead of a specific risk measure, we define the risk-averse bandit problem with arbitrary (but fixed) risk measure and the corresponding regret. We focus on risk measures defined as a function of the first two moments (the mean and the variance). This generalizes the setting of [16] from linear to arbitrary functions, while considering notion of regret similar to [10].

We present two motivating examples of our framework: (1) We consider the *threshold variance* problem, where we have the usual bandit setting and interested in the means of the distributions (of the arms), but would like to chose only from those arms that has the variance smaller than a specified threshold. One possible formalization of this problem leads us to the risk-averse regret minimization with discontinuous function of the mean and the variance used as a risk measure. (2) Consider a risk measure that is a linear combination of the mean and the square root of the variance, where both the summands are of the same order. This is a natural variant of the mean-variance optimization and is a continuous function of the mean and the variance.

Our main results are as follows: (1) First we present an algorithm, namely,  $\varphi$ -LCB, which belongs to the wide family of Lower (Upper) Confidence Bound algorithm (the descendants of UCB algorithm of [3], see also, e.g. [2], [9]), and prove logarithmic risk-averse regret bounds for all continuous functions. (2) Second, we present an example of a discontinuous function where no natural algorithm (based on the optimism in face of uncertainty principle) can achieve sublinear regret. (3) Finally, we present another algorithm, namely,  $\varphi$ -LCB2, that makes learning feasible with the mild assumption that no arm hits the discontinuity points. Our proof approach is similar to [16] and [10], while the latter used slightly different KL-divergence based version of the algorithm ([11]).

*Other related works.* In the bandit setting risk-aversion has been approached from different perspectives. [8] designs an algorithm that uses conditional value at risk (CVaR) as a risk measure. However, they aim at minimizing the usual expected regret under the assumption that the best mean arm is also the best risk-aversion arm, which is completely different from our goal. [22] derive PAC-bounds on the single- and multi-period risk for several different risk measures, nevertheless, the PAC-style of their results makes it inapplicable to our problem. [15] considers the deviations of the regret in the standard setting, which seem to address the same issues, but it remains unclear if their results can be connected to risk-averse regret minimization.

*Organization.* In Section 2 we introduce the notations to be used, formally state the problem, and present some examples which can be modeled in our framework.

In Sections 3.1 and 3.2 we discuss two cases of the main problem and present the corresponding algorithms together with the risk-averse regret bounds. Section 4 discusses open problems and the possible extensions of the setting. The paper concludes with the proofs of the main theorems in Section 5.

## 2 The problem

Let  $\mathcal{L}_2$  denote the set of distributions supported on  $[0, 1]$ . We consider the stochastic multi-armed bandit setting with  $K$  arms and  $\nu_1, \dots, \nu_K \in \mathcal{L}_2$  being the distributions of arms. At time step  $t$  the learner chooses arm  $a_t$  to pull and receives a sample  $X_{a_t, T_{a_t}(t)}$  drawn from  $\nu_{a_t}$ , where  $T_i(t)$  is the number of times that arm  $i$  is pulled by the  $t$ -th time step, that is,

$$T_i(t) = \sum_{s=1}^t \mathbb{I}[a_s = i] \quad .$$

We consider the case where the learner is given a risk measure  $R : \mathcal{L}_2 \rightarrow \mathbb{R}$ . The risk measure of arm  $i$  is  $R_i = R(\nu_i)$ . This measure defines the best arm  $i^*$  by

$$i^* = \underset{i=1..K}{\operatorname{argmin}} R_i$$

and the goal of the algorithm is to identify that arm. The performance of the algorithm is measured by means of risk-averse regret:

$$\mathcal{R}_n = \sum_{t=1}^n R_{a_t} - \sum_{t=1}^n R_{i^*} = \sum_{t=1}^n R_{a_t} - n \cdot R_{i^*} \quad .$$

Note that this corresponds to the notion of pseudo-regret for stochastic bandits, but there is no regret notion in our setting that directly corresponds to true regret in stochastic bandits. One could try to define true regret as the difference of risk measures applied to the empirical distributions of the algorithm and the best arm (similar to [16]). However, then the algorithm could be punished even for switching between the best arms, which can be an undesirable feature.

Some examples of such risk measures are  $R(X) = \mathbb{E}[X]$  with  $X$  being a random variable (usual stochastic bandit) and  $R(X) = \frac{1}{\lambda} \log \mathbb{E}[\exp \lambda X]$ , considered in [10].

In this paper we focus on the risk measures of the following form:

$$R(X) = f(\mathbb{E}[X], \operatorname{Var}(X)) \quad .$$

In other words, the learner is supplied by a function  $f : D \rightarrow \mathbb{R}$ , where  $D = [0, 1] \times [0, 1]^1$ . If we denote the risk measure of arm  $i$  by  $f_i$ , i.e.  $f_i = f(\mu_i, \sigma_i^2)$ ,

---

<sup>1</sup> The domain of the second argument can be restricted to  $[0, \frac{1}{4}]$ , since for a random variable which takes values in  $[0, 1]$ , the variance is upper bounded by  $\frac{1}{4}$ .

where  $\mu_i$  and  $\sigma_i^2$  are the mean and the variance of the  $i$ -th arm respectively, then  $i^* = \operatorname{argmin}_{i=1..K} f_i$  and the regret is

$$\mathcal{R}_n = \sum_{t=1}^n f_{a_t} - \sum_{t=1}^n f_{i^*} = \sum_{t=1}^n f_{a_t} - n \cdot f_{i^*} .$$

Our class of risk measures is rich enough to model a lot of interesting problems:

1. **Standard Bandit:**  $f(x, y) = x$ . This is the standard stochastic multi-armed bandit setting.
2. **Variance Minimization:**  $f(x, y) = y$ . This is the variance minimization problem, considered in [21].
3. **Mean-variance Bandit:**  $f(x, y) = x + \lambda \cdot y$ . This is a version of the problem considered in [16]. A related and natural variant is  $f(x, y) = x + \lambda \sqrt{y}$ , where both summands are of the same order.
4. **Threshold Variance:**  $f(x, y) = x \mathbb{I}[y < v] + \mathbb{I}[y \geq v]$ . This risk measure can be used to model threshold variance problem described in Section 1.
5. **Log-Exponential Risk:**  $f(x, y) = x + \frac{\lambda}{2}x^2 + \frac{\lambda}{2}y$ . This measure can be seen as an approximation to the coherent risk measure, considered in [10]:  $\frac{1}{\lambda} \log \mathbb{E}[\exp \lambda X]$ , when it is restricted to the first two moments.

Our goal is to study conditions on the function  $f$  under which learning is possible.

### 3 Our Results

We distinguish between two cases of the problem: continuous and discontinuous functions  $f$ . In the continuous case we prove that learning is possible for every function. In the discontinuous case we present an example where learning is not possible. The negative example motivates a restriction, and we show that under the restriction learning is feasible.

#### 3.1 Continuous functions

In this section we will show that learning is possible for any continuous function  $f$ . We start with a characterization of continuous functions that will be used to present the algorithm.

**Lemma 1.** *For every continuous function  $f : D \rightarrow \mathbb{R}$ , there exists a function  $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , such that*

1.  $\varphi(0) = 0$ ;
2.  $\varphi$  is a strictly increasing function;
3.  $|f(\mathbf{x}_2) - f(\mathbf{x}_1)| \leq \varphi(\|\mathbf{x}_2 - \mathbf{x}_1\|_1)$  for all  $\mathbf{x}_1, \mathbf{x}_2 \in D$ .

As an example, consider an  $\alpha$ -Hölder continuous function  $f$ : in this case  $\varphi(z) = cz^\alpha$  would satisfy the conditions of Lemma 1 by the definition of  $\alpha$ -Hölder continuity. But Lemma 1 is stated for every continuous function: as another example, consider the continuous function

$$h(x) = \begin{cases} \frac{-1}{\ln(x/2)} & \text{if } x \in D \text{ and } x > 0 \\ 0 & \text{if } x = 0 \end{cases} . \quad (1)$$

It is not  $\alpha$ -Hölder continuous for any  $\alpha$ , but  $\varphi(z) = h(z)$  satisfies the conditions of Lemma 1 for  $f(x, y) = h(x)$ .

We will use Lemma 1 to construct a high-confidence interval for  $f$  from the confidence intervals for its arguments. We start by defining the empirical mean and the empirical variance of arm  $i$ :

$$\hat{\mu}_{i,t} = \frac{1}{t} \sum_{s=1}^t X_{i,s} \quad \text{and} \quad \hat{\sigma}_{i,t}^2 = \frac{1}{t} \sum_{s=1}^t (X_{i,s} - \hat{\mu}_{i,t})^2 .$$

The following concentration results are the basis for our argument.

**Lemma 2 (Chernoff-Hoeffding bound).** *For every  $i = 1, \dots, K$ ,  $t = 1, \dots, n$ , and  $\delta \in (0, \frac{1}{2})$ , with probability at least  $1 - 2\delta$*

$$|\hat{\mu}_{i,t} - \mu_i| \leq \sqrt{\frac{\ln \frac{1}{\delta}}{2t}} .$$

**Lemma 3 (Lemma 2 from [1]).** *For all  $i = 1, \dots, K$ ,  $t = 1, \dots, n$ , and  $\delta \in (0, \frac{1}{4Kn})$ , with probability at least  $1 - 4Kn\delta$*

$$|\hat{\sigma}_{i,t}^2 - \sigma_i^2| \leq 5\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} . \quad (2)$$

From Lemma 1, Lemma 2, and Lemma 3 we can construct the following high-confidence bound for  $f$ :

$$|f(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2) - f_i| \leq \varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \right) . \quad (3)$$

The algorithm  $\varphi$ -LCB will at time step  $t$  choose an arm that minimizes the corresponding lower confidence bound:

$$a_t = \operatorname{argmin}_{i=1..K} \left[ f(\hat{\mu}_{i,T_i(t-1)}, \hat{\sigma}_{i,T_i(t-1)}^2) - \varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i(t-1)}} \right) \right] . \quad (4)$$

The algorithm chooses arm  $i$  if  $f(\hat{\mu}_{i,T_i(t-1)}, \hat{\sigma}_{i,T_i(t-1)}^2)$  is really small or if  $\varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i(t-1)}} \right)$  is big. The former means that the algorithm tries to exploit

the arm that has small estimated risk measures, while the latter means that the estimate for the arm  $i$  is rough and the algorithm tries to improve it by exploring this arm further. In other words, the  $\varphi$ -LCB algorithm tries to deal with exploration-exploitation trade-off using the so-called optimism in face of uncertainty principle.

**Parameters:** Confidence level  $\delta$ ;  
**For all time steps**  $t = 1, 2, \dots, n$ , **repeat**

1. Compute  $a_t = \operatorname{argmin}_{i=1..K} \left[ f(\hat{\mu}_{i, T_i(t-1)}, \hat{\sigma}_{i, T_i(t-1)}^2) - \varphi \left( 6 \sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i(t-1)}} \right) \right]$ .
2. Output  $a_t$  as a decision.
3. Receive  $X_{a_t, T_{a_t}(t)} \sim \nu_{a_t}$ .

**Fig. 1.** The  $\varphi$ -LCB algorithm

Theorem 1 states the regret bound of the  $\varphi$ -LCB algorithm.

**Theorem 1 (Feasibility of learning).** *Consider a continuous function  $f$ , then for  $\delta \in (0, \frac{1}{4Kn})$  with probability at least  $1 - 4Kn\delta$  the regret of the  $\varphi$ -LCB algorithm at time  $n$  is upper bounded by:*

$$\mathcal{R}_n \leq \sum_{i: \Delta_i > 0} \frac{18 \cdot \Delta_i \cdot \ln \frac{1}{\delta}}{(\varphi^{-1}(\Delta_i/2))^2} + \sum_{i: \Delta_i > 0} \Delta_i ,$$

where  $\Delta_i = f_i - f_{i^*}$ . Moreover, for  $n > 4K$ , if the algorithm is run with  $\delta = \frac{1}{n^2}$ , then with probability at least  $1 - \frac{4K}{n}$  the regret is upper bounded by:

$$\mathcal{R}_n \leq \sum_{i: \Delta_i > 0} \frac{36 \cdot \Delta_i}{(\varphi^{-1}(\Delta_i/2))^2} \ln n + \sum_{i: \Delta_i > 0} \Delta_i .$$

**Efficiency.** Theorem 1 shows that learning is feasible for every continuous function. We now discuss the efficiency of the algorithm with respect to different classes of continuous functions.

1. **Lipschitz functions:** when  $f$  is  $L$ -Lipschitz, i.e.  $\varphi(z) = Lz$ , the regret bound is

$$\mathcal{R}_n \leq \sum_{i: \Delta_i > 0} \frac{144 \cdot L^2}{\Delta_i} \ln n + \sum_{i: \Delta_i > 0} \Delta_i$$

and the dependence on  $\Delta_i$  in front of  $\ln n$  matches the dependence in the regret of the  $\varphi$ -LCB algorithm in the standard stochastic bandit problem. The worse constant  $(144L^2)$  term is an artifact of doing such general analysis. This case covers the standard bandit and the variance minimization problems with  $L = 1$ , the log-exponential risk problem with  $L = 1 + \lambda$ , and the mean-variance bandit problem with  $f(x, y) = x + \lambda y$  in which  $L = \max\{1, \lambda\}$ .

2. **Hölder functions:** when  $f$  is  $\alpha$ -Hölder continuous, i.e.  $\varphi(z) = Lz^\alpha$ , the regret bound is

$$\mathcal{R}_n \leq \sum_{i:\Delta_i>0} \frac{36 \cdot (2 \cdot L)^{\frac{2}{\alpha}}}{(\Delta_i)^{\frac{2-\alpha}{\alpha}}} \ln n + \sum_{i:\Delta_i>0} \Delta_i .$$

This case covers the mean-variance problem with  $f(x, y) = x + \lambda\sqrt{y}$  which is  $\frac{1}{2}$ -Hölder continuous with  $L = \max\{1, \lambda\}$ . Note that the dependence on  $\Delta_i$  in this case is worse than for Lipschitz functions, but it is still polynomial.

3. **Non-Hölder functions:** to demonstrate how efficiency can decrease for the general class of continuous functions, consider  $f(x, y) = h(x)$  from (1), then  $\varphi(z) = h(z)$  and the regret bound becomes

$$\mathcal{R}_n \leq \sum_{i:\Delta_i>0} 9 \cdot \Delta_i \cdot e^{4/\Delta_i} \ln n + \sum_{i:\Delta_i>0} \Delta_i .$$

We can see that the term in front of  $\ln n$  grows exponentially as  $\Delta_i$  goes to 0 in comparison to the polynomial growth for Lipschitz and Hölder functions.

**Remark 1** *Note that it is possible to design an anytime version of  $\varphi$ -LCB for the case when  $n$  is not known in advance. To do so, at each time step we take  $\delta = \varepsilon_t$ , where  $\varepsilon_t$  is a sequence decreasing at an appropriate rate. However, we do not pursue this direction further.*

### 3.2 Discontinuous functions

The case of discontinuous functions is more tricky. We present a negative example and a partially positive result. We start with an example of a discontinuous function  $f$  where no algorithm following the optimism in face of uncertainty principle can achieve sublinear regret.

*Example 1.* Consider the following discontinuous function: Let

$$f(x, y) = \begin{cases} 1 & \text{if } x = 0.5 \text{ and } y = 0.1; \\ \frac{1}{2} & \text{if } y \geq 0.5; \\ 0 & \text{otherwise .} \end{cases}$$

Consider two arms 1 and 2 such that  $\mu_1 = 0.5$  and  $\sigma_1^2 = 0.1$  and  $\mu_2 = 1$  and  $\sigma_2^2 = 0.75$ . Then any algorithm based on the optimism in face of uncertainty principle will keep on choosing arm 1 with non-negligible probability. This is because if the estimate of the algorithm is not precisely the discontinuity point, then arm 1 will be chosen due to optimism.

However, in the case when no arm hits the discontinuity point, learning is possible as we will show. Let  $d_i(x, y) = |x - \mu_i| + |y - \sigma_i^2|$  be the distance to the point representing  $i$ -th arm. Define  $\Omega_f$  to be the set of discontinuities of  $f$  and  $d_\Omega(x, y) = \inf_{(z_1, z_2) \in \Omega_f} \{|z_1 - x| + |z_2 - y|\}$  to be the distance to the closest discontinuity point. We will show that learning is possible under the following assumption.

**Assumption 1** For each arm  $i$  there exists  $\varepsilon > 0$  such that  $f$  is continuous in  $B_i(\varepsilon) = \{(x, y) \in D : d_i(x, y) \leq \varepsilon\}$ .

Let us introduce  $e_i = \sup\{\varepsilon > 0 : f \text{ is continuous in } B_i(\varepsilon)\} = d_\Omega(\mu_i, \sigma_i^2)$ , then by Lemma 1 there exists a function  $\varphi_i$  that satisfies the required condition, but only in  $B_i(e_i)$  instead of  $D$ . So when our estimated values are in  $B_i(e_i)$  we can use the same algorithm as before. We present a new algorithm  $\varphi$ -LCB2 that first pulls each arm some amount of times, such that with high probability  $(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2)$  is in  $B_i(e_i)$  for each arm, in other words, that  $d_i(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2) \leq e_i$ . If we would know  $e_i$  in advance, then to ensure this condition with high probability it is enough (from Lemma 2 and Lemma 3) that

$$6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \leq e_i .$$

Hence, we would need to pull each arm  $18e_i^{-2} \ln \frac{1}{\delta}$  times. But since  $e_i$  is not known in advance, we would pull each arm until its distance to  $(\mu_i, \sigma_i)$  is twice less than distance to the closest discontinuity point. Formally, the algorithm chooses each arm until

$$d_i(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2) \leq \frac{1}{2} d_\Omega(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2) . \quad (5)$$

At the time when this happens, we can be sure that  $(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2) \in B_i(e_i)$  and this procedure does not increase the number of pulls too much. To ensure (5) with high probability it is enough that

$$6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \leq \frac{1}{2} d_\Omega(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2) . \quad (6)$$

After ensuring this for each arm, the algorithm proceeds as the  $\varphi$ -LCB algorithm, but uses  $\varphi_i$  for each arm instead of a common function  $\varphi$ :

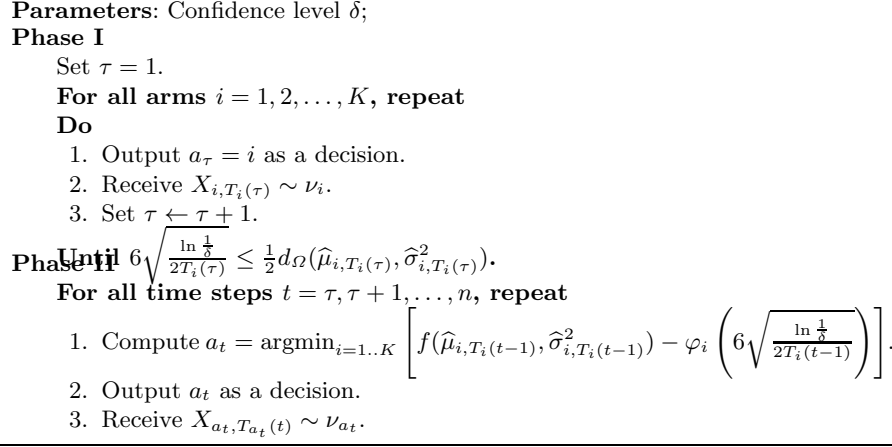
$$a_t = \operatorname{argmin}_{i=1..K} \left[ f(\hat{\mu}_{i,T_i(t-1)}, \hat{\sigma}_{i,T_i(t-1)}^2) - \varphi_i \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i(t-1)}} \right) \right] . \quad (7)$$

Note that constructing  $\varphi_i$  requires knowledge of  $e_i$ , but this can also be avoided if we construct it in the estimated (and smaller) region, defined at the time, when (6) occurs. The following theorem states the regret bound of the resulting algorithm.

**Theorem 2.** Consider function  $f$  that satisfies Assumption 1. Then for  $\delta \in (0, \frac{1}{4Kn})$  with probability at least  $1 - 4Kn\delta$  for all  $n \geq \sum_{i=1..K} 162 \cdot e_i^{-2} \ln \frac{1}{\delta}$  the regret of the  $\varphi$ -LCB2 algorithm at time  $n$  is upper bounded by:

$$\mathcal{R}_n \leq \sum_{i:\Delta_i > 0} \Delta_i \left( 162e_i^{-2} + \frac{18}{(\varphi_i^{-1}(\Delta_i/2))^2} \right) \ln \frac{1}{\delta} + \sum_{i:\Delta_i > 0} \Delta_i$$





**Fig. 2.**  $\varphi$ -LCB2 algorithm

where  $\Delta_i$  and  $e_i$  as defined before. Moreover, if the algorithm is run with  $\delta = \frac{1}{n^2}$ , then with probability at least  $1 - \frac{4K}{n}$  for all  $n \geq \sum_{i=1..K} 324 \cdot e_i^{-2} \ln n$  the regret is upper bounded by:

$$\mathcal{R}_n \leq \sum_{i:\Delta_i>0} \Delta_i \left( 324e_i^{-2} + \frac{36}{(\varphi_i^{-1}(\Delta_i/2))^2} \right) \ln n + \sum_{i:\Delta_i>0} \Delta_i .$$

The theorem can be applied to our motivating example: the threshold variance problem. There are two continuous regions, when  $y < v$  and when  $y \geq v$ . In either case we can take  $\varphi(z) = z$  (in fact, we can take any increasing function for the region  $y \geq v$ , since  $f$  is just a constant there) and then the bound becomes

$$\mathcal{R}_n \leq \sum_{i:\Delta_i>0} 4 \left( 81 \cdot e_i^{-2} \Delta_i + \frac{36}{\Delta_i} \right) \ln n + \sum_{i:\Delta_i>0} \Delta_i .$$

Actually, in this case the bound can be improved, since after Phase I the algorithm would know which arms have variance greater than  $v$  and it would not pull them at all. Hence, for such arms term  $4\frac{36}{\Delta_i} \ln n$  can be removed. Note that the efficiency of the algorithm depends on how fast we can compute  $d_\Omega(\hat{\mu}_{i,T_i(\tau)}, \hat{\sigma}_{i,T_i(\tau)}^2)$ : For the threshold variance problem it can be done efficiently, because  $d_\Omega(x, y) = |y - v|$ , i.e. it can be done in constant time.

## 4 Conclusion and discussion

We described a framework for the risk-averse regret minimization without restriction to any particular risk measure. For a specific class of risk measures, which are functions of the mean and the variance, we proposed two algorithms that achieve logarithmic regret: one for the case of continuous functions and the

one for the case of discontinuous functions. In the former case we proved logarithmic regret bound for any continuous function, while in the latter the problem need to satisfy a mild and reasonable assumption that arms should not hit the discontinuity points of the risk measure. Under this condition, the algorithms presented achieves the logarithmic regret.

We believe that assumption 1 might not be a necessary condition for learning. For example, even for the case when the risk measure is the Dirichlet function of the mean (which is continuous nowhere), it maybe be possible to design a sound algorithm, following the lines of [5].

We remark that achieving optimal constants was not our goal and it is very likely that our bounds can be improved. An open problem, which we have not addressed in our work, is lower bounds on the risk-averse regret. Since the standard bandit problem is a particular case of our problem, we know that in this case the bound is tight (up to a constant), but obtaining a general lower bound remains an interesting research direction. Another open problem is the extension of our results to other classes of functions. While a long-term goal would be to consider general functionals, the class of coherent risk measures could be a plausible next step. It is interesting to note that while classes of coherent risk measures and general functions of the mean and the variance intersect, there is no inclusion in either direction. Finally, it is an interesting question to consider the best arm identification problem (e.g. [4]) in the context of our framework. This problem is usually referred to as a pure exploration problem, where the goal is to explore the arms in the most efficient way, focusing on minimizing the notion of simple regret.

## 5 Proofs

*Proof (Lemma 1).* We will prove the lemma by directly constructing a candidate function, satisfying the stated conditions. First note that by Heine-Cantor theorem  $f$  is uniformly continuous, since the domain  $D$  is compact. Consider a sequence  $\varepsilon_i = 2^{-i}$  for  $i \geq 0$ , then for every such  $\varepsilon_i$  there exists  $\delta_i > 0$ , such that  $\|\mathbf{x}_2 - \mathbf{x}_1\|_1 < \delta_i \Rightarrow |f(\mathbf{x}_2) - f(\mathbf{x}_1)| < \varepsilon_i$  by uniform continuity. We now decrease each  $\delta_i$  such that  $\delta_i \leq \varepsilon_i$  (if it is not the case). This does not invalidate the previous implication. Afterwards we construct the function  $\psi$ . First,  $\psi(0) = 0$ . Then for any  $z < \delta_0$  we define

$$k(z) = \max \{i : z \leq \delta_i\} .$$

Then  $\psi(z) = \varepsilon_{k(z)} = 2^{-k(z)}$  for  $z < \delta_0$ . Now we need to deal with the case when  $z \geq \delta_0$ . For this note that the fact  $\|\mathbf{x}_2 - \mathbf{x}_1\|_1 < \delta \Rightarrow |f(\mathbf{x}_2) - f(\mathbf{x}_1)| < \varepsilon$  for any  $\mathbf{x}_1, \mathbf{x}_2 \in D$  implies  $\|\mathbf{x}_2 - \mathbf{x}_1\|_1 < 2\delta \Rightarrow |f(\mathbf{x}_2) - f(\mathbf{x}_1)| < 2\varepsilon$  for any  $\mathbf{x}_1, \mathbf{x}_2 \in D$ . To see this, assume the former is true and fix  $\mathbf{x}_1, \mathbf{x}_2$  such that  $\|\mathbf{x}_2 - \mathbf{x}_1\|_1 < 2\delta$ . Take  $\mathbf{z} = \frac{1}{2} \cdot (\mathbf{x}_2 + \mathbf{x}_1)$ , then for both  $\mathbf{x}_1$  and  $\mathbf{x}_2$ :  $\|\mathbf{x}_i - \mathbf{z}\|_1 < \delta$  and hence  $|f(\mathbf{x}_i) - f(\mathbf{z})| < \varepsilon$ . But then

$$|f(\mathbf{x}_2) - f(\mathbf{x}_1)| \leq |f(\mathbf{x}_2) - f(\mathbf{z})| + |f(\mathbf{z}) - f(\mathbf{x}_1)| < 2\varepsilon .$$

We use the just proven fact to define  $\psi$  for  $z \geq \delta_i$ . Let  $i$  be the smallest  $i$  such that  $z < 2^i \delta_0$ , then  $\psi(z) = 2^i \varepsilon_0$ . To unify both cases we introduce

$$a_i = \begin{cases} \delta_{-i} & \text{if } i \leq 0 \\ 2^i \delta_0 & \text{if } i > 0 \end{cases}.$$

Letting  $k(z) = \min\{i : z \leq a_i\}$ , for  $z > 0$ . We then have that  $\psi(z) = 2^{k(z)}$ . By construction,  $\psi$  satisfy Condition 1 and Condition 3 of the lemma (for any  $\mathbf{x}_2, \mathbf{x}_1 \in D : \|\mathbf{x}_2 - \mathbf{x}_1\|_1 \leq a_{k(\|\mathbf{x}_2 - \mathbf{x}_1\|_1)}$ , and then  $|f(\mathbf{x}_2) - f(\mathbf{x}_1)| \leq 2^{k(\|\mathbf{x}_2 - \mathbf{x}_1\|_1)} = \psi(\|\mathbf{x}_2 - \mathbf{x}_1\|_1)$ ). Also,  $\psi$  is well-defined, since for all  $z > 0$  (1) there exists some  $i$  such that  $z \leq 2^i \delta_0$ ; and (2) we have that  $\forall i : \delta_i \leq \epsilon_i = 2^{-i}$  and thus  $k(z) \geq -i$  for  $2^{-i} \leq z$ . To deal with Condition 2, we can take any strictly increasing function  $\varphi$  that dominates  $\psi$  at every point. For example, we can linearly interpolate between discontinuity points, i.e. define  $\varphi$  as

$$\varphi(z) = \frac{1}{a_{k(z)} - a_{k(z)-1}} \left( 2^{k(z)-1}(z - a_{k(z)-1}) + 2^{k(z)}(a_{k(z)} - z) \right)$$

for  $z > 0$  and  $\varphi(0) = 0$ . It is strictly increasing (because  $\psi$  is increasing, which we get from the definition of  $k(z)$ ) and Condition 3 follows from  $\psi(z) \leq \varphi(z)$  for  $z \geq 0$ .

*Proof (Theorem 1).* The proof is similar to Theorem 1 from [16] with minor modifications. We start with the following standard regret decomposition (recall that  $\Delta_i = f_i - f_{i^*}$ ).

$$\mathcal{R}_n = \sum_{t=1}^n f_{a_t} - \sum_{t=1}^n f_{i^*} = \sum_{i: \Delta_i > 0} \Delta_i T_i(n) \quad (8)$$

Hence, our task is reduced to bounding  $T_i(n)$  for each arm. First, let  $\mu_i^{(2)}$  be the second moment of the distribution of the arm  $i$ , i.e.  $\mu_i^{(2)} = \mathbb{E}[Y^2]$ , where  $Y \sim \nu_i$ . Then

$$\hat{\mu}_{i,t}^{(2)} = \frac{1}{t} \sum_{s=1}^t X_{i,s}^2$$

is the estimator of  $\mu_i^{(2)}$ . Now we define a high probability event

$$A = \left\{ \forall t = 1, \dots, n; \forall i = 1, \dots, K : |\hat{\mu}_{i,t} - \mu_i| \leq \sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \text{ and } |\hat{\mu}_{i,t}^{(2)} - \mu_i^{(2)}| \leq \sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \right\}. \quad (9)$$

Using Lemma 2 and union bound, one can get that  $\mathbb{P}[A^c] \leq 4Kn\delta$ . From Lemma 2 in [1], we get that (2) holds on  $A$  and, consequently, (3) also holds on  $A$  (for every  $t = 1, \dots, n$  and  $i = 1, \dots, K$ ).

Now let us consider the moment when arm  $i$  is chosen at some time step  $t$ . It means that its lower confidence index was lower than that of the best arm (by (4)):

$$\begin{aligned} f(\hat{\mu}_{i, T_i(t-1)}, \hat{\sigma}_{i, T_i(t-1)}^2) - \varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i(t-1)}} \right) &\leq \\ f(\hat{\mu}_{i^*, T_i^*(t-1)}, \hat{\sigma}_{i^*, T_i^*(t-1)}^2) - \varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i^*(t-1)}} \right) &. \end{aligned}$$

We also know that on the event  $A$  (by (3)):

$$f_i - \varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i(t-1)}} \right) \leq f(\hat{\mu}_{i, T_i(t-1)}, \hat{\sigma}_{i, T_i(t-1)}^2)$$

and

$$f(\hat{\mu}_{i^*, T_i^*(t-1)}, \hat{\sigma}_{i^*, T_i^*(t-1)}^2) - \varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i^*(t-1)}} \right) \leq f_{i^*} .$$

Combining the last three inequalities,

$$f_i - 2\varphi \left( 6\sqrt{\frac{\ln \frac{1}{\delta}}{2 \cdot T_i(t-1)}} \right) \leq f_{i^*} .$$

Since  $\varphi$  is strictly increasing function it has a well-defined inverse  $\varphi^{-1}$  and we can bound  $T_i(t-1)$  as follows:

$$T_i(t-1) \leq \frac{18 \cdot \ln \frac{1}{\delta}}{(\varphi^{-1}(\Delta_i/2))^2} .$$

If  $t$  is the last time when arm  $i$  is pulled, then  $T_i(n) = T_i(t-1) + 1$  and hence

$$T_i(n) \leq \frac{18 \cdot \ln \frac{1}{\delta}}{(\varphi^{-1}(\Delta_i/2))^2} + 1 . \quad (10)$$

Inserting this into (8) gives us the stated regret bound.

*Proof (Theorem 2).* Again, as in Theorem 1, we are going to use regret decomposition (8). Hence, we will focus on bounding  $T_i(n)$  for each arm  $i$ . We define the event  $A$  as in (9) and everything we are deriving next is conditioned on  $A$ . We introduce the following stopping times  $\lambda_i$  as

$$\lambda_i = \inf \left\{ t : 6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \leq \frac{1}{2} \cdot d_\Omega(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2) \right\} .$$

Then we have

$$T_i(n) = \lambda_i + \tilde{T}_i(n) ,$$

where  $\tilde{T}_i(n)$  is the number of times the arm  $i$  was pulled during the second phase of the algorithm. Conditioned on  $A$  it can be bounded as in Theorem 1 by (10) with corresponding  $\varphi_i$ . Next we focus on  $\lambda_i$ . If we define

$$\tilde{\lambda}_i = \inf \left\{ t : 6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \leq \frac{e_i}{3} \right\} = \inf \left\{ t : 6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} \leq \frac{d_\Omega(\mu_i, \sigma_i^2)}{3} \right\} ,$$

then, at time  $\tilde{\lambda}_i$  Condition (6) is necessarily fulfilled:

$$\begin{aligned} 6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} &\leq \frac{d_\Omega(\mu_i, \sigma_i^2)}{3} \\ &\leq \frac{d_i(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2)}{3} + \frac{d_\Omega(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2)}{3} \\ &\leq \frac{1}{3} \cdot 6\sqrt{\frac{\ln \frac{1}{\delta}}{2t}} + \frac{d_\Omega(\hat{\mu}_{i,t}, \hat{\sigma}_{i,t}^2)}{3} . \end{aligned}$$

Hence  $\lambda_i \leq \tilde{\lambda}_i = 162 \cdot e_i^{-2} \ln \frac{1}{\delta}$ . Combining this together with (10) and (8) gives the stated result.

## References

1. Antos, A., Grover, V., Szepesvári, C.: Active learning in heteroscedastic noise. *Theoretical Computer Science* 411(29), 2712–2728 (2010)
2. Audibert, J.Y., Munos, R., Szepesvári, C.: Tuning bandit algorithms in stochastic environments. In: *Algorithmic Learning Theory*. pp. 150–165. Springer (2007)
3. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3), 235–256 (2002)
4. Bubeck, S., Munos, R., Stoltz, G.: Pure exploration in multi-armed bandits problems. In: *Algorithmic Learning Theory*. pp. 23–37. Springer (2009)
5. Cover, T.M.: On determining the irrationality of the mean of a random variable. *The Annals of Statistics* pp. 862–871 (1973)
6. Defourny, B., Ernst, D., Wehenkel, L.: Risk-aware decision making and dynamic programming. In: *Selected for oral presentation at the NIPS-08 Workshop on Model Uncertainty and Risk in Reinforcement Learning*, Whistler, Canada (2008)
7. Even-Dar, E., Kearns, M., Wortman, J.: Risk-sensitive online learning. In: *Algorithmic Learning Theory*. pp. 199–213. Springer (2006)
8. Galichet, N., Sebag, M., Teytaud, O.: Exploration vs exploitation vs safety: Risk-averse multi-armed bandits. *JMLR: Workshop and Conference Proceedings* 29 pp. 245–260 (2013)
9. Garivier, A., Cappé, O.: The kl-ucb algorithm for bounded stochastic bandits and beyond. *JMLR: Workshop and Conference Proceedings* 19 pp. 359–376 (2011)
10. Maillard, O.A.: Robust risk-averse stochastic multi-armed bandits. In: *Algorithmic Learning Theory, Lecture Notes in Computer Science*, vol. 8139, pp. 218–233 (2013)

11. Maillard, O.A., Munos, R., Stoltz, G., et al.: A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In: 24th Annual Conference on Learning Theory: COLT'11 (2011)
12. Markowitz, H.: Portfolio selection. *The journal of finance* 7(1), 77–91 (1952)
13. Patek, S.D.: On terminating markov decision processes with a risk-averse objective function. *Automatica* 37(9), 1379–1386 (2001)
14. Rockafellar, R.T.: Coherent approaches to risk in optimization under uncertainty. *Tutorials in operations research, INFORMS* (2007)
15. Salomon, A., Audibert, J.Y.: Deviations of stochastic bandit regret. In: *Algorithmic Learning Theory*. pp. 159–173. Springer (2011)
16. Sani, A., Lazaric, A., Munos, R.: Risk-aversion in multi-armed bandits. In: *Advances in Neural Information Processing Systems* 25. pp. 3284–3292 (2012)
17. Sharpe, W.F.: Mutual fund performance. *The Journal of Business* 39(1), 119–138 (1966)
18. Shen, Y., Stannat, W., Obermayer, K.: Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization* 51(5), 3652–3672 (2013)
19. Shen, Y., Tobia, M.J., Sommer, T., Obermayer, K.: Risk-sensitive reinforcement learning. *Neural Computation* (2014)
20. Von Neumann, J., Morgenstern, O.: *The theory of games and economic behavior* (1947)
21. Warmuth, M.K., Kuzmin, D.: Online variance minimization. In: *Learning Theory*, pp. 514–528. Springer (2006)
22. Yu, J.Y., Nikolova, E.: Sample complexity of risk-averse bandit-arm selection. In: *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. pp. 2576–2582. AAAI Press (2013)