PAC Bandits with Risk Constraints

Yahel David
Technion, Haifa, Israel
Proteantecs Ltd, Haifa, Israel
yahel83@gmail.com

Balázs Szörényi
szorenyi.balazs@gmail.com

Mohammad Ghavamzadeh
Google DeepMind, CA
ghavamza@google.com

Shie Mannor
Technion, Haifa, Israel
shie@ee.technion.ac.il

Nahum Shimkin
Technion, Haifa, Israel
shimkin@ee.technion.ac.il

Abstract

We study the problem of best arm identification with risk constraints within the setting of fixed confidence pure exploration bandits (PAC bandits). The goal is to stop as fast as possible, and with high confidence return an arm whose mean is ϵ-close to the best arm among those that satisfy a risk constraint, namely their α-quantile functions are larger than a threshold β. For this risk-sensitive bandit problem, we propose an algorithm and prove an upper-bound on its sample complexity for the general case of sub-Gaussian arms' distributions. We also prove a lower-bound for this general case that shows our derived upper-bound is near-optimal (up to logarithmic factors). Both our upper and lower bounds have similar form to the risk-neutral PAC bandits results of (Even-Dar *et al.* 2006) and (Mannor and Tsitsiklis 2004), respectively. We also prove a lower-bound for our problem when the arms' distributions are Gaussian, which is smaller than our general lower-bound, but is stronger in the sense that it applies to any instance of the (Gaussian) problem. This lower-bound is in terms of the KL divergence and has similar behavior to the risk-neutral PAC bandits results of (Kaufmann *et al.* 2016).

1 Introduction

In *stochastic multi-armed bandit* problem, a learner interacts with a number of (possibly infinite) unknown distributions (each corresponds to an arm) by selecting an arm and sampling from its distribution (pulling an arm) at each round. There have been two main formulations of this problem: *cumulative regret* and *pure exploration*. In *cumulative regret*, which is the classic formulation, the goal of the learner is to come up with a strategy to pull the arms in a way to maximize the sum of its (expected) rewards, or in other words, to minimize its (pseudo) regret (Robbins 1952; Thompson 1933; 1935). In the *pure exploration* formulation, the learner is not evaluated while interacting with the arms, its performance is only measured when it stops and answers a pre-defined question about the arms, such as which arm is the best (has the highest mean). The problem of best arm identification, which is probably the most popular pure exploration bandit problem (e.g., (Even-Dar *et al.* 2006; Audibert *et al.* 2010; Kalyanakrishnan *et al.* 2012; Gabillon *et al.* 2012; Kaufmann and Kalyanakrishnan 2013; Kaufmann *et al.* 2016)), has been studied in two different settings: **1)** *fixed budget* in which the objective is to find an arm whose mean is ϵ-close to the best arm with the largest possible confidence $1-\delta$, using a fixed budget of rounds, and **2)** *fixed confidence* or *PAC bandits* in which the goal is to stop as soon as possible and return an arm whose mean is ϵ-close to the best arm with a fixed confidence $1-\delta$. In this paper, we consider the *fixed confidence pure exploration* setting, where the goal is to return an arm with the highest mean among those that satisfy a risk constraint (are not too risky).

Throughout the long and active history of the stochastic multi-armed bandit problem, its different formulations and settings have provided powerful tools in various domains including finance, energy management, online marketing, and robotics. However, most of the developed algorithms in this area only focus on the mean rewards of the arms and ignore the risk imposed by the variation in these random variables, which can be costly in many applications. For example, consider an arm whose distribution has *extremely high* reward with probability 0.1 and *low* reward with probability 0.9. Although this arm might have the highest mean among the arms, its obtained reward is low most of the time, which may not be acceptable in many applications (an investor may go bankrupt before this distribution hits the jackpot and returns the extremely high reward). This is why bandit researchers have recently begun to study the so called *risk-averse* bandits (e.g., (Sani *et al.* 2012; Maillard 2013; Yu and Nikolova 2013; Vakili and Zhao 2016)). These papers aim at finding an arm with the optimal risk measure (instead of the one with the highest mean) in some form, such as mean-variance (Markowitz 1952), value at risk (VaR) (Artzner *et al.* 1999), expected shortfall (Rockafellar and Uryasev 2000), or a dynamic entropic risk measure (Maillard 2013).

In this paper, we study risk averse fixed-confidence pure exploration bandit, where our goal is "not" to return an arm with an optimal risk measure, but instead is to find an arm with the highest mean (similar to the classic best arm identification setting) among those that satisfy a risk constraint, namely their α-quantile functions are larger than a threshold β (see Section 2 for the detailed description of our setting). Our objective is most closely related to that in (Galichet *et al.* 2013; Zimin *et al.* 2014; Haskell *et al.* 2016), although it is important to note that all these papers study the cumulative regret formulation. (Galichet *et al.* 2013) utilizes the strong assumption that the arm with the highest mean

has the lowest risk. This assumption significantly simplifies the problem, but can be unrealistic in many scenarios. In contrast, we derive bounds for the general case, without utilizing such assumption. (Zimin *et al.* 2014) considers a general setting for risk-constrained multi-armed bandit optimization, where the objective function might depend on both the means and variances of the distributions. They propose an algorithm and prove an upper-bound for its regret. Somewhat similar, but distribution independent (i.e., worst-case) results are also obtained in (Haskell *et al.* 2016). Our setting can be cast into this framework for some special cases, like the Gaussian bandits. However, the upper-bounds in these papers depend on some parameters that are not clear how to be computed in specific cases, and no matching lower-bound has been presented, which leaves the tightness of the results unclear. In contrast, our results are shown to be tight in some specific cases and have clear interpretation. The contributions of this paper can be summarized as

1) We propose an algorithm for our risk constraint bandit problem and prove an upper-bound on its sample complexity (Section 4). The algorithm is for the general case of sub-Gaussian arms' distributions. Our upper-bound depends mainly on **(a)** Δ_k^*, i.e., the gap between the mean of arm k and the optimal arm and **(b)** $\Delta_{\beta,k}$, i.e., a term that quantifies how far the risk of arm k is from the threshold β (we properly define these terms in Section 2). This upper-bound has a form similar to the classic result on risk-neutral PAC bandits (Even-Dar *et al.* 2006), and in fact, reproduces it when the risk constraints are removed.

2) We prove a lower-bound on the sample complexity of our risk constraint bandit problem (in its general case of sub-Gaussian arms' distributions) that shows our derived upper-bound is near-optimal (Section 3.1). More importantly, our lower-bound has similar form as the existing results in risk-neutral PAC bandits (Mannor and Tsitsiklis 2004).

3) We derive a lower-bound on the sample complexity of our risk constraint bandit problem for the special case of Gaussian arms' distributions (Section 3.2). This lower-bound is expressed in terms of the Kullback-Leibler divergence and in one hand is smaller than our general lower-bound (Section 3.1), and on the other hand, is stronger in the sense that it applies to *any* instance of the (Gaussian) problem. Although we do not show a matching upper-bound for this setting, similar to our general lower-bound, our Gaussian result has similar behavior (up to the risk-related parameters) to those in risk-neutral PAC bandits (Kaufmann *et al.* 2016).

2 Problem Formulation

We consider a finite set of K arms $\mathcal{K} = \{1, \dots, K\}$, where each arm $k \in \mathcal{K}$ is characterized by a distribution ν_k (either bounded, e.g., in $[0, 1]$, or sub-Gaussian) with unknown mean μ_k and CDF F_k. We measure the risk of each arm by its quantile function defined as

Definition 1. *For an arm $k \in \mathcal{K}$, we define its quantile function for a probability $0 < \alpha < 1$ as*

$$\rho_k(\alpha) \triangleq \inf\big\{x \in \mathbb{R} \mid 1 - \alpha \le F_k(x)\big\}.$$

We denote by $\mathcal{K}_{\beta,0} \triangleq \big\{k \in \mathcal{K} \mid \rho_k(\alpha) \ge \beta\big\}$ and $\mathcal{K}_{\beta,\epsilon_\rho} \triangleq \big\{k \in \mathcal{K} \mid \rho_k(\alpha - \epsilon_\rho) \ge \beta\big\}$ the set of arms whose quantile functions for levels α and $(\alpha - \epsilon_\rho)$ are larger than a threshold β, respectively. We call these two sets, the set of *feasible* arms and the set of *ϵ_ρ-approximately feasible* arms, respectively. We also denote by $\mu^* \triangleq \max_{k \in \mathcal{K}_{\beta,0}} \mu_k$ and $\mu' \triangleq \max_{k \in \mathcal{K}_{\beta,\epsilon_\rho}} \mu_k$ the highest means in these two sets. Note that since whenever $\rho(\alpha) \ge \beta$ then $\rho(\alpha - \epsilon_\rho) \ge \beta$, we have $\mathcal{K}_{\beta,0} \subseteq \mathcal{K}_{\beta,\epsilon_\rho}$ and $\mu^* \le \mu'$. Finally, we define $\mathcal{K}^* \triangleq \big\{k \in \mathcal{K}_{\beta,\epsilon_\rho} \mid \mu_k \ge \mu^* - \epsilon_\mu\big\}$ as the set of *ϵ_ρ-approximately feasible and ϵ_μ-approximately optimal* arms. We define our problem of finding the best low-risk arm as the problem of returning an arm that belongs to the desirable set $\mathcal{K}^*$.

Our problem can be formalized as a game between a stochastic bandit environment and a learner. Before the game begins, the learner is given the risk level α, the risk threshold β, the accuracy parameters ϵ_ρ and ϵ_μ, and the confidence parameter δ. At each round $t = 1, 2, \dots$, the learner pulls an arm $k_t \in \mathcal{K}$ and observes a reward sampled from its distribution ν_{k_t}. The rewards obtained from each arm are i.i.d. samples from its distribution. The goal is that the learner stops in a finite number of rounds T and returns an arm k_T that with probability at least $(1 - \delta)$ belongs to $\mathcal{K}^*$, i.e., $\mathbb{P}(k_T \notin \mathcal{K}^*) \le \delta$. We call such a learner $(\epsilon_\mu, \epsilon_\rho, \delta)$-*correct*. The performance of the learner is then measured by the number of rounds T either in expectation or with high probability.

Finally, for each arm $k \in \mathcal{K}$, we define the following notions of gap that we will use them throughout the paper:

$$\Delta_k^* \triangleq \max(0, \max_{k' \in \mathcal{K}_{\beta,0}} \mu_{k'} - \mu_k) \triangleq \max(0, \mu^* - \mu_k),$$

$$\Delta_k' \triangleq \max(0, \max_{k' \in \mathcal{K}_{\beta,\epsilon_\rho}} \mu_{k'} - \mu_k) \triangleq \max(0, \mu' - \mu_k),$$

and

$$\Delta_{k,\beta} \triangleq \sup\big\{F_k(x) \mid x \le \beta\big\} - (1-\alpha) = F_k(\beta) - (1-\alpha). \tag{1}$$

3 Lower Bounds

In this section, we present our lower-bound results for the general case of sub-Gaussian and the special case of Gaussian arms' distributions in Sections 3.1 and 3.2, respectively.

3.1 Lower-Bound for the General Case of Sub-Gaussian Arms

In this section, we consider the general case that the distributions of the arms are sub-Gaussian. The sample complexity lower-bound result is based on several hand-crafted problem instances, and has a form similar to the classic result on risk-neutral PAC bandits (Mannor and Tsitsiklis 2004).

Theorem 2. *Fix an algorithm and the parameters $0 < \epsilon_\mu < 1/32$, $0 < \epsilon_\rho < 1/16$, $7/8 < \alpha < 1$, $1/16 < \beta < 1/8$, and $0 < \delta < 0.01$. If the expected number of samples used by*

the algorithm is

$$\mathbb{E}[T] \leq \min_{k' \in \mathcal{K}} \sum_{k \in \{\mathcal{K} \setminus k'\}} \frac{\ln\left(1/9\delta^{-1}\right)}{900 \max\left((5\epsilon_\mu + 4\Delta'_k)^2, (16\Delta_{k,\beta})^2\right)} - K \tag{2}$$

for any set of sub-Gaussian[1] *arms $\mathcal{K}$, then the algorithm is not $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct.*

Proof. In this proof we show that if for a given algorithm the bound in Equation (2) is violated for a specific set of arms, then, that algorithm can't be $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct for another set of arms. We use the notation $f_k(\cdot)$ for the probability density function (p.d.f.) of arm $k \in \mathcal{K}$.

Let's assume that the algorithm is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct and its sample complexity violates the bound in Equation (2) for some hypothesis H_0 for which the arms p.d.f. satisfy Equation (3) for some constants $1/4 < a_k < 3/4$ and $1/4 < b_k < 3/4$.

$$f_k(x) = \begin{cases} 0 & x < 0 \\ a_k & 0 \leq x < 1/8 \\ 2 - a_k & 1/8 \leq x < 1/4 \\ b_k & 1/4 \leq x < 1/2 \\ 2 - b_k & 1/2 \leq x < 3/4 \\ 2 - a_k & 3/4 \leq x < 7/8 \\ a_k & 7/8 \leq x < 1 \\ 0 & x \geq 1 \end{cases}. \tag{3}$$

Note that by Equation (3) it follows that the constant a_k determines whether the arm is too risky or not, and the constant b_k determines the mean reward of the arm.

Now, we define the following set of hypotheses $\{H_1, ..., H_K\}$, where $f_l^{H_k}(x)$ stands for the p.d.f. of arm l under hypothesis k. For $k = 1, .., K$, we define H_k as follows, for $l \neq k$, H_k coincides with the true distribution, namely,

$$f_l^{H_k}(x) = f_l(x), \quad l \neq k.$$

For $l = k$, we construct $f_k^{H_k}(x)$ as follows

$$f_k^{H_k}(x) = \begin{cases} 0 & x < 0 \\ a_k^{H_k} & 0 \leq x < 1/8 \\ 2 - a_k^{H_k} & 1/8 \leq x < 1/4 \\ b_k^{H_k} & 1/4 \leq x < 1/2 \\ 2 - b_k^{H_k} & 1/2 \leq x < 3/4 \\ 2 - a_k^{H_k} & 3/4 \leq x < 7/8 \\ a_k^{H_k} & 7/8 \leq x < 1 \\ 0 & x \geq 1 \end{cases}. \tag{4}$$

where $a_k^{H_k} = \min\left(a_k, (1-\alpha)/\beta\right)$,and $b_k^{H_k} = \min\left(b_k, \min_{k' \in \mathcal{K}_{\beta,\epsilon_\rho}} b_{k'} - 5\epsilon_\mu\right)$.

Note that under hypothesis H_k, arm k is the only correct arm. Therefore, under hypothesis $H_k, k = 1, .., K$, an $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct algorithm should return arm k with probability larger than $1 - \delta$. We use $\mathbb{E}_k^H$ and $\mathbb{P}_k^H$ to denote the expectation and probability, respectively, under the algorithm being considered and hypothesis H_k. For every $k \in \mathcal{K}$ let T_k stand for the number of samples from arm k and let

$$t'_k = \left\lfloor \frac{\ln\left(\frac{1}{9\delta}\right)}{900\left(\max\left(\left(b_k - b_k^{H_k}\right)^2, \left(a_k - a_k^{H_k}\right)^2\right)\right)} \right\rfloor,$$

As explained before, we suppose that the algorithm is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct under H_0, and that Equation (2) is violated, so, $\mathbb{E}_0^H[T_k] < t'_k$ for some $k \in \mathcal{K}$. We will show that this algorithm cannot be $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct under hypothesis H_k. Therefore, there are some sets of arms for which every $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct algorithm must have $\mathbb{E}_0^H[T_k] \geq t'_k$ for all $k \in \mathcal{K}$.

Define the following events , for $k \in \mathcal{K}$:

- $A'_k = \{T_k \leq 4t'_k\}$. It easily follows from $4t'_k\left(1 - \mathbb{P}_0^H(A'_k)\right) \leq \mathbb{E}_0^H[T_k]$ that if $\mathbb{E}_0^H[T_k] \leq t'_k$, then $\mathbb{P}_0^H(A'_k) \geq \frac{3}{4}$.
- Let B'_k stand for the event under which the chosen arm at termination is k, and B'^C_k for its complement. Since $\mathbb{P}_0^H\left(B'_{k'}\right) > \frac{1}{2}$ can hold for one arm at most, it follows that $(\exists k')(\forall k \neq k')\mathbb{P}_0^H\left(B'^C_k\right) > \frac{1}{2}$,
- Let $x_i^k(a)$, $x_i^k(\overline{a})$, $x_i^k(b)$ and $x_i^k(\overline{b})$ be the following R.V.

$$\begin{aligned} x_i^k(a) &= \begin{cases} 1 & f_k(x_i^k) = a_k \\ 0 & f_k(x_i^k) \neq a_k \end{cases} \\ x_i^k(\overline{a}) &= \begin{cases} 1 & f_k(x_i^k) = 2 - a_k \\ 0 & f_k(x_i^k) \neq 2 - a_k \end{cases} \\ x_i^k(b) &= \begin{cases} 1 & f_k(x_i^k) = b_k \\ 0 & f_k(x_i^k) \neq b_k \end{cases} \\ x_i^k(\overline{b}) &= \begin{cases} 1 & f_k(x_i^k) = 2 - b_k \\ 0 & f_k(x_i^k) \neq 2 - b_k \end{cases} \end{aligned} \tag{5}$$

  where x_i^k is the i-th sample from arm k.
- Let $C_k(a)$, $C_k(\overline{a})$, $C_k(b)$, $C_k(\overline{b})$, be events under which for every number of samples $t \leq 4t_k$ obtained from arm k, the sums $\sum_{i \leq t} x_i^k(a)$, $\sum_{i \leq t} x_i^k(\overline{a})$, $\sum_{i \leq t} x_i^k(b)$, $\sum_{i \leq t} x_i^k(\overline{b})$ are bounded as follows

$$C_k(a) \triangleq \left\{ \max_{1 \leq t \leq 4t'_k} \sum_{i \leq t} \left[x_i^k(a) - \tfrac{a_k}{4}\right] < \phi_k a_k \right\}$$

$$C_k(\overline{a}) \triangleq \left\{ \max_{1 \leq t \leq 4t'_k} \sum_{i \leq t} \left[x_i^k(\overline{a}) - \tfrac{2-a_k}{4}\right] < \phi_k (2 - a_k) \right\}$$

$$C_k(b) \triangleq \left\{ \max_{1 \leq t \leq 4t'_k} \sum_{i \leq t} \left[x_i^k(b) - \tfrac{b_k}{4}\right] < \phi_k b_k \right\}$$

$$C_k(\overline{b}) \triangleq \left\{ \max_{1 \leq t \leq 4t'_k} \sum_{i \leq t} \left[x_i^k(\overline{b}) - \tfrac{2-b_k}{4}\right] < \phi_k (2 - b_k) \right\}$$

[1] All the arm distributions in the proof have support $[0, 1]$. Nonetheless, the argument can be easily extended to sub-Gaussian arms.

where $\phi_k = 8\sqrt{2}\sqrt{t'_k}$. Now, by using Kolmogorov's inequality we bound $\mathbb{P}_0^H(C_k(a))$, $\mathbb{P}_0^H(C_k(\overline{a}))$, $\mathbb{P}_0^H(C_k(b))$ and $\mathbb{P}_0^H(C_k(\overline{b}))$. Kolmogorov's inequality states that the sum $S_t = \sum_{i=1}^{t} z_i$ of zero-mean iid random variables (z_i) satisfies $\mathbb{P}(\max_{1\le t\le n}|S_t| \ge a) \le \frac{\text{Var}[S_n]}{a^2}$ (Theorem 22.4, in p. 287 of (Billingsley 1995)). By applying it to the RVs

$$\begin{aligned} y_i^k(a) &= x_i^k(a) - a_k/4 \\ y_i^k(\overline{a}) &= x_i^k(\overline{a}) - (2-a_k)/4 \\ y_i^k(b) &= x_i^k(b) - b_k/4 \\ y_i^k(\overline{b}) &= x_i^k(\overline{b}) - (2-b_k)/4 \end{aligned}$$

and by using the fact that the variance of the sum of n iid Bernoulli R.V.s with parameter p is $np(1-p)$, we obtain that

$$\begin{aligned} \mathbb{P}_0^H\left(C_k^C(a)\right) &\le \frac{\text{Var}(\sum_{i=1}^{4t'_k} y_i^k(a))}{\left(8\sqrt{2}a_k\sqrt{t'_k}\right)^2} \le \tfrac{1}{32}\,, \\ \mathbb{P}_0^H\left(C_k^C(\overline{a})\right) &\le \frac{\text{Var}(\sum_{i=1}^{4t'_k} y_i^k(\overline{a}))}{\left(8\sqrt{2}(2-a_k)\sqrt{t'_k}\right)^2} \le \tfrac{1}{32}\,, \\ \mathbb{P}_0^H\left(C_k^C(b)\right) &\le \frac{\text{Var}(\sum_{i=1}^{4t'_k} y_i^k(b))}{\left(8\sqrt{2}b_k\sqrt{t'_k}\right)^2} \le \tfrac{1}{32}\,, \\ \mathbb{P}_0^H\left(C_k^C(\overline{b})\right) &\le \frac{\text{Var}(\sum_{i=1}^{4t'_k} y_i^k(\overline{b}))}{\left(8\sqrt{2}(2-b_k)\sqrt{t'_k}\right)^2} \le \tfrac{1}{32}\,, \end{aligned} \tag{6}$$

where $C_k^C(\cdot)$ is the complementary of $C_k(\cdot)$. So, $\mathbb{P}_0^H(C_k(\cdot)) \ge \frac{31}{32}$ for a, $\overline{a}$, b and $\overline{b}$. Hence, $\mathbb{P}_0^H(C'_k) \ge \frac{7}{8}$, where $C'_k = C_k(a) \cap C_k(\overline{a}) \cap C_k(b) \cap C_k(\overline{b})$.

Define now the intersection event $S'_k = A'_k \cap B_k'^C \cap C'_k$. We have just shown that for every $k \ne k'$ it holds that $\mathbb{P}_0^H(A'_k) \ge \frac{3}{4}$, $\mathbb{P}_0^H(B_k'^C) > \frac{1}{2}$, $\mathbb{P}_0^H(C'_k) \ge \frac{7}{8}$, from which it follows that

$$\mathbb{P}_0^H\left(S_k\right) > \frac{1}{8} \quad \text{for} \quad k \ne k'. \tag{7}$$

Now, we let h be the history of the process (the sequence of chosen arms and obtained rewards). For a given history, at time t', for every $k \in \mathcal{K}$, the probability of choosing the next arm is the same under H_0 and under H_k. Also, by the hypotheses definition, the reward probability is the same, unless the chosen arm is k. Furthermore, as $\mathbb{P}_k^H$ is absolutely continuous w.r.t. $\mathbb{P}_0^H$, for any history h, by their Radon-Nikodym derivative and by the definition of the hypotheses it follows that

$$\frac{d\mathbb{P}_k^H}{d\mathbb{P}_0^H}(h) \ge R(a)R(\overline{a})R(b)R(\overline{b}), \tag{8}$$

where

$$\begin{aligned} R(a) &= \left(a_k^{H_k}/a_k\right)^{\sum_{t=1}^{T_k(h)} x_i^k(a)}\,, \\ R(\overline{a}) &= \left(2-a_k^{H_k}/2-a_k\right)^{\sum_{t=1}^{T_k(h)} x_i^k(\overline{a})}\,, \\ R(b) &= \left(b_k^{H_k}/b_k\right)^{\sum_{t=1}^{T_k(h)} x_i^k(b)}\,, \\ R(\overline{b}) &= \left(2-b_k^{H_k}/2-b_k\right)^{\sum_{t=1}^{T_k(h)} x_i^k(\overline{b})}\,. \end{aligned}$$

Now, when the events S'_k holds, it is obtained that

$$\begin{aligned} R(a) &\ge \left(1-\frac{a_k-a_k^{H_k}}{a_k}\right)^{(a_k/4)t'_k+8\sqrt{2}a_k\sqrt{t'_k}}\,, \\ R(\overline{a}) &\ge \left(1+\frac{a_k-a_k^{H_k}}{2-a_k}\right)^{((2-a_k)/4)t'_k-8\sqrt{2}(2-a_k)\sqrt{t'_k}}\,, \\ R(b) &\ge \left(1-\frac{b_k-b_k^{H_k}}{b_k}\right)^{(b_k/4)t'_k+8\sqrt{2}b_k\sqrt{t'_k}}\,, \\ R(\overline{b}) &\ge \left(1+\frac{b_k-b_k^{H_k}}{2-b_k}\right)^{((2-b_k)/4)t'_k-8\sqrt{2}(2-b_k)\sqrt{t'_k}}\,. \end{aligned} \tag{9}$$

Then, by the fact that for $\epsilon > 0$ it follows that $(1-\epsilon)^{\frac{1}{\epsilon}} \ge e^{-1}$ and that $(1+\epsilon)^{\frac{1}{\epsilon}} \ge e^{1-\epsilon}$, by Equation (9) and the definition of t'_k it is obtained that

$$\begin{aligned} R(a)R(\overline{a}) &\ge e^{\frac{\ln(9\delta)}{3600(2-a_k)} - \frac{16\sqrt{2}\sqrt{-\ln(9\delta)}}{30}}, \quad h \in S'_k\,, \\ R(b)R(\overline{b}) &\ge e^{\frac{\ln(9\delta)}{3600(2-b_k)} - \frac{16\sqrt{2}\sqrt{-\ln(9\delta)}}{30}}, \quad h \in S'_k. \end{aligned} \tag{10}$$

Also, $-\ln(9\delta) \ge 1.55\sqrt{-\ln(9\delta)}$, so,

$$R(a)R(\overline{a})R(b)R(\overline{b}) \ge 9\delta, \quad h \in S'_k\,. \tag{11}$$

Therefore, we obtain the following inequalities,

$$\begin{aligned} \mathbb{P}_k^H\left(B_k'^C\right) \ge \mathbb{P}_k^H\left(S'_k\right) &= \mathbb{E}_0^H\left[\frac{d\mathbb{P}_k^H}{d\mathbb{P}_0^H}(h) I_{\left\{h\in S'_k\right\}}\right] \\ &\ge 9\delta\mathbb{P}_0^H\left(S'_k\right) \\ &\ge \frac{9}{8}\delta > \delta, \quad \forall k \ne k', \end{aligned}$$

where the last inequality follows from (7).

We found that if an algorithm is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct under hypothesis H_0 and $\mathbb{E}_0[T_k] \le t'_k$ for some $k \ne k'$, then, under hypothesis H_k this algorithm returns an arm that is either not ϵ_μ or not satisfies the risk condition with probability of δ or more, hence the algorithm is not $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct. Therefore, any $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct algorithm must satisfy $\mathbb{E}_0[T_k] \ge t'_k$ for all of arms except possibly for one (namely, for the one k' for which $\mathbb{P}_0\left(B_{k'}'^C\right) \le \frac{1}{2}$, if such k' exists).

Now, since under H_0, it follows that $F_k(\beta) = a_k\beta$ it is obtained that

$$a_k - a_k^{H_k} = \frac{1}{\beta}\max\left(\Delta_{k,\beta}, 0\right)\,. \tag{12}$$

Furthermore, as

$$\Delta'_k = \frac{1}{4}\left(b_k - \min\left(b_k, \min_{k'\in\mathcal{K}_{\beta,\epsilon_\rho}} b_{k'}\right)\right)\,,$$

it easily obtained that

$$b_k - b_k^{H_k} \le 5\epsilon_\mu + 4\Delta'_k\,, \tag{13}$$

where $\Delta'_k = \max\Big(0, \max_{k' \in \mathcal{K}_{\beta,\epsilon_\rho}} \mu_{k'} - \mu_k\Big)$.

Hence, by substituting Equations (12) and (13) in the definition of t'_k, Equation (2) is obtained. □

3.2 Lower-Bound for Gaussian Arms

Our next lower-bound result is for the scenario in which all the arms have normal distributions, i.e., $\nu_k = \mathcal{N}(\mu_k, \sigma_k^2)$, $\forall k \in \mathcal{K}$. Our lower-bound behaves similarly (up to the risk-related parameters) to the lower-bound for the unconstrained PAC bandits (Kaufmann *et al.* 2016).

Theorem 3. *Let all the arms have Gaussian distributions and $\delta \leq 0.04$. Then, for every $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct algorithm, we have*

$$\mathbb{E}[T] \geq \min_{k' \in \mathcal{K}} \sum_{k \in \mathcal{K} \setminus k'} \frac{\ln\left(\frac{1}{9\delta}\right)}{22^2 \eta_k}, \tag{14}$$

where $\eta_k \triangleq \mathrm{kl}(\nu_k, \nu_{H_k})$ is the Kullback-Leibler divergence from ν_{H_k} to ν_k, the distribution $\nu_{H_k} = \mathcal{N}(\mu_{H_k}, \sigma_{H_k}^2)$ is Gaussian with mean $\mu_{H_k} = \max(\mu_k, 2\epsilon_\mu + \mu')$ and standard deviation $\sigma_{H_k} = \min(\sigma_k, \widetilde{\sigma}_k)$, where $\widetilde{\sigma}_k$ is the maximum σ for which $\rho(\alpha) \geq \beta$, for the Gaussian distribution $\mathcal{N}(\mu_{H_k}, \sigma^2)$.

Proof. We begin the proof by recalling that the kl-divergence of two Gaussians may be written as

$$\eta_k = \mathrm{kl}(\nu_k, \nu_{H_k}) = \ln\left(\frac{\sigma_{H_k}}{\sigma_k}\right) + \frac{\Delta_k^2 + \sigma_k^2 - \sigma_{H_k}^2}{2\sigma_{H_k}^2},$$

where $\Delta_k = \mu_{H_k} - \mu_k$. We now define the following set of hypotheses $\{H_0, H_1, \ldots, H_K\}$ and denote by $\mu_l^{H_k}$ and $\sigma_l^{H_k}$ the mean and standard deviation of arm l under hypothesis H_k. The hypothesis H_0 is the true hypothesis, i.e., $\mu_k^{H_0} = \mu_k$ and $\sigma_k^{H_0} = \sigma_k$, $\forall k \in \mathcal{K}$. Each hypothesis H_k, $k \in \mathcal{K}$, is defined as: for each arm $l \neq k$, its mean and standard deviation are the true ones, i.e., $\mu_l^{H_k} = \mu_l$ and $\sigma_l^{H_k} = \sigma_l$, $l \neq k$, and the mean and standard deviation of arm k, denoted by μ_{H_k} and σ_{H_k}, are defined as in the statement of Theorem 3.

We have defined the set of hypotheses $\{H_i\}_{i=1}^K$ in a way that under each hypothesis H_k in this set, arm k be the unique arm in $\mathcal{K}_{H_k}^*$, i.e., the set of ϵ_ρ-approximately feasible and ϵ_μ-approximately optimal arms. Thus, under each hypothesis H_k, any $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct algorithm should return arm k with probability at least $1 - \delta$. To see this, we should show that for any arm $l \neq k$, if $l \in \mathcal{K}_{\beta,\epsilon_\rho}^{H_k}$, then $\mu_l < \mu_{H_k}^* - \epsilon_\mu$ to make sure that $l \notin \mathcal{K}_{H_k}^*$.

$$\mu_{H_k}^* - \epsilon_\mu \overset{\text{(a)}}{=} \mu_{H_k} - \epsilon_\mu \overset{\text{(b)}}{\geq} \mu'_{H_0} + 2\epsilon_\mu - \epsilon_\mu \overset{\text{(c)}}{\geq} \mu_l + \epsilon_\mu > \mu_l,$$

(a) from the definition of μ_{H_k}, we have $\mu_{H_k} \geq \mu'_{H_0} + 2\epsilon_\mu$, and since $\mu'_{H_0} \geq \mu^*_{H_0}$, we have $\mu^*_{H_k} = \max(\mu^*_{H_0}, \mu_{H_k}) = \mu_{H_k}$,
(b) comes from $\mu_{H_k} \geq \mu'_{H_0} + 2\epsilon_\mu$,
(c) follows from the fact that if $l \in \mathcal{K}_{\beta,\epsilon_\rho}^{H_k}$, then $l \in \mathcal{K}_{\beta,\epsilon_\rho}^{H_0}$, and thus, $\mu_l \leq \mu'_{H_0}$.

For each $k \in \mathcal{K}$, we denote by T_k the number of rounds at which arm k has been pulled and define $t_k = \lfloor \frac{\ln\left(\frac{1}{9\delta}\right)}{22^2 \eta_k} \rfloor$. Now suppose that an algorithm is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct under hypothesis H_0 and $\mathbb{E}_{H_0}[T_k] \leq t_k$ for some $k \in \mathcal{K}$. We will prove that this algorithm cannot be $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct under hypothesis H_k. Therefore, an $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct algorithm must have $\mathbb{E}_{H_0}[T_k] > t_k$ for all $k \in \mathcal{K}$.

To prove the above statement, we define the following events for the arm k for which $\mathbb{E}_{H_0}[T_k] \leq t_k$.

1) $\mathcal{E}_{1,k} = \{T_k \leq 4t_k\}$. It is easy to show that $4t_k\big(1 - \mathbb{P}_{H_0}(\mathcal{E}_{1,k})\big) \leq \mathbb{E}_{H_0}[T_k]$, which together with $\mathbb{E}_{H_0}[T_k] \leq t_k$, gives us $\mathbb{P}_{H_0}(\mathcal{E}_{1,k}) \geq \frac{3}{4}$.

2) $\mathcal{E}_{2,k}$ is the event that the arm returned by the algorithm is k. We denote by $\mathcal{E}_{2,k}^C$ the complement of this event. Since $\mathbb{P}_{H_0}(\mathcal{E}_{2,k'}) > \frac{1}{2}$ can hold for at most one arm, we may write

$$\exists k' : \mathbb{P}_{H_0}(\mathcal{E}_{2,k}^C) > \frac{1}{2},\ \forall k \neq k'.$$

3) $\mathcal{E}_{3,k}$ is the event that for any number of samples $1 \leq t \leq 4t_k$ obtained from arm k, we have the following bound on the sum $\sum_{i=1}^t \zeta_k(x_{k,i})$,

$$\mathcal{E}_{3,k} \triangleq \Big\{ \max_{1 \leq t \leq 4t_k} \sum_{i=1}^{t} \big(\zeta_k(x_{k,i}) - \eta_k\big) < 21\sqrt{t_k \max\left(\eta_k, \eta_k^2\right)}\Big\},$$

where $x_{k,i}$ is the value of the i'th sample from arm k and[2]

$$\zeta_k(x) \triangleq \ln\left(\frac{\sigma_{H_k}}{\sigma_k}\right) + \frac{(x - \mu_{H_k})^2}{2\sigma_{H_k}^2} - \frac{(x - \mu_k)^2}{2\sigma_k^2}.$$

Kolmogorov's inequality states that the sum of zero-mean i.i.d. random variables $\{z_i\}_{i=1}^n$, i.e., $S_t = \sum_{i=1}^t z_i$, satisfies $\mathbb{P}\big(\max_{1 \leq t \leq n} |S_t| \geq a\big) \leq \frac{\mathrm{Var}[S_n]}{a^2}$ (Theorem 22.4, pp. 287 of (Billingsley 1995)). Applying the Kolmogorov's inequality when random variables $\{z_{k,i}\}_{i=1}^{4t_k}$ are defined as $z_{k,i} = \zeta(x_{k,i}) - \eta_k$, we obtain[3]

$$\begin{aligned}
\mathbb{P}_{H_0}(\mathcal{E}_{3,k}^C) &\leq \frac{\mathrm{Var}\big[\sum_{i=1}^{4t_k} z_{k,i}\big]}{\big(21\sqrt{t_k \max\left(\eta_k, \eta_k^2\right)}\big)^2} \\
&\leq \frac{4t_k \times 13 \max\left(\eta_k, \eta_k^2\right)}{\big(21\sqrt{t_k \max\left(\eta_k, \eta_k^2\right)}\big)^2} \qquad (15) \\
&< \frac{1}{8}, \qquad (16)
\end{aligned}$$

where (15) comes from the following lemma whose proof is omitted due to space constraints.

Lemma 4. *Under the conditions of Eq.* (16)*, we have* $\mathrm{Var}\big[\sum_{i=1}^{4t_k} z_{k,i}\big] \leq 4t_k \times 13 \max(\eta_k, \eta_k^2)$.

We showed above that under our assumptions, for any arm $k \neq k'$, we have $\mathbb{P}_{H_0}(\mathcal{E}_{1,k}) \geq \frac{3}{4}$, $\mathbb{P}_{H_0}(\mathcal{E}_{2,k}^C) > \frac{1}{2}$, and $\mathbb{P}_{H_0}(\mathcal{E}_{3,k}) \geq \frac{7}{8}$. Thus, for the intersection of these events,

[2] Note that $\zeta_k(x) = \ln\big(d\mathbb{P}_{H_k}(x)/d\mathbb{P}_{H_0}(x)\big)$.

[3] It is easy to show that the random variables $\{z_{k,i}\}_{i=1}^{4t_k}$ are zero mean, i.e., $\mathbb{E}_{H_0}[z_{k,i}] = 0$, and i.i.d.

i.e., $\mathcal{E}_k = \mathcal{E}_{1,k} \cap \mathcal{E}_{2,k}^C \cap \mathcal{E}_{3,k}$, we may write $\mathbb{P}_{H_0}(\mathcal{E}_k) > 1/8$, for $k \neq k'$, because

$$\begin{aligned}
\mathbb{P}_{H_0}(\mathcal{E}_k) &= \mathbb{P}_{H_0}(\mathcal{E}_{1,k} \cap \mathcal{E}_{2,k}^C \cap \mathcal{E}_{3,k}) \\
&= 1 - \mathbb{P}_{H_0}(\mathcal{E}_{1,k}^C \cap \mathcal{E}_{2,k} \cap \mathcal{E}_{3,k}^C) \\
&\geq 1 - \left[\mathbb{P}_{H_0}(\mathcal{E}_{1,k}^C) + \mathbb{P}_{H_0}(\mathcal{E}_{2,k}) + \mathbb{P}_{H_0}(\mathcal{E}_{3,k}^C)\right] \\
&= 1 - \left[3 - \mathbb{P}_{H_0}(\mathcal{E}_{1,k}) - \mathbb{P}_{H_0}(\mathcal{E}_{2,k}^C) - \mathbb{P}_{H_0}(\mathcal{E}_{3,k})\right] \\
&> \frac{3}{4} + \frac{1}{2} + \frac{7}{8} - 2 = \frac{1}{8}. \qquad (17)
\end{aligned}$$

Let h_t be the history of the process (the sequence of the selected arms and observed rewards) up to and not including time t. Note that for a given history h_t, the probability of choosing an arm $k \in \mathcal{K}$ at time t is the same under the hypotheses H_0 and H_k. Moreover, from the definition of these hypotheses, the reward probability at time t is the same, unless arm k is selected, i.e., $k_t = k$. Finally, since all the arm distributions are Gaussian, $\mathbb{P}_{H_k}$ is absolutely continuous w.r.t. $\mathbb{P}_{H_0}$, and thus, for any history h, we may write their Radon-Nikodym derivative as

$$\frac{d\mathbb{P}_{H_k}}{d\mathbb{P}_{H_0}}(h) \geq e^{-\sum_{t=1}^{T_k(h)} \zeta_k(x_{k,t})}, \qquad (18)$$

where $T_k(h)$ is the number of times arm k selected in h. On the intersection event $\mathcal{E}_k$, we may write

$$\begin{aligned}
\sum_{t=1}^{T_k(h)} \zeta_k(x_{k,t}) &< 4t_k\eta_k + 21\sqrt{t_k \max(\eta_k, \eta_k^2)} \qquad (19) \\
&\leq 4\frac{\ln(1/9\delta)}{22^2} + 21\sqrt{\frac{\ln(1/9\delta)}{22^2\eta_k}\max(\eta_k, \eta_k^2)} \qquad (20) \\
&= 4\frac{\ln(1/9\delta)}{22^2} + \frac{21}{22}\sqrt{\ln(1/9\delta)} \qquad (21) \\
&\leq \ln(1/9\delta), \qquad (22)
\end{aligned}$$

(19) comes from the definition of $\mathcal{E}_{3,k}$ and the fact that on $\mathcal{E}_{1,k}$, we have $T_k(h) \leq 4t_k$; (20) follows from the definition of t_k and under the assumption that $\eta_k \leq \frac{\ln(1/9\delta)}{22^2}$. Note that $t_k = 0$, for $\eta_k > \frac{\ln(1/9\delta)}{22^2}$; (21) is under the assumption that $\max(\eta_k, \eta_k^2) = \eta_k$, which means $\eta_k \leq 1$; (22) follows from $\sqrt{\ln(1/9\delta)} \leq \ln(1/9\delta)$, which holds for $\delta \leq 0.04$.

From Eq. (22), on event $\mathcal{E}_k$, we have $\frac{d\mathbb{P}_{H_k}}{d\mathbb{P}_{H_0}}(h) \geq 9\delta$. Therefore, we may write the following inequalities:

$$\begin{aligned}
\mathbb{P}_{H_k}(\mathcal{E}_{2,k}^C) &\geq \mathbb{P}_{H_k}(\mathcal{E}_k) \\
&= \mathbb{E}_{H_0}\left[\frac{d\mathbb{P}_{H_k}}{d\mathbb{P}_{H_0}}(h) I_{\{h \in \mathcal{E}_k\}}\right] \\
&\geq 9\delta \times \mathbb{P}_{H_0}(\mathcal{E}_k) \\
&> \frac{9}{8}\delta \qquad (23) \\
&> \delta \qquad (24)
\end{aligned}$$

for all $k \neq k'$, where (23) follows from Eq. (17). Recall that under hypothesis H_k, an algorithm is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct, if and only if it returns arm k. Therefore, what Eq. (24) shows is that if an algorithm is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct under hypothesis H_0 and $\mathbb{E}_{H_0}[T_k] \leq t_k$ for some arm $k \neq k'$, then under hypothesis H_k this algorithm does not return arm k with probability larger than δ, and thus, is not $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct. As a result, any $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct algorithm must satisfy $\mathbb{E}_{H_0}[T_k] > t_k$ for all the arms, except possibly for one, namely arm k' for which $\mathbb{P}_{H_0}(\mathcal{E}_{2,k}^C) \leq \frac{1}{2}$. Thus, Eq. (14) follows. □

4 Algorithm

In this section, we first propose an algorithm for the general case of sub-Gaussian arms; see Algorithm 1 for the pseudocode. It adapts the arm-elimination principle to the risk-averse setting. At each round, it pulls the most promising arm, and keeps track of both its mean and the relevant quantile estimates. These values are used to discard an arm when it turns out to be sub-optimal or violating the risk constraint, and to decide if an arm is a good solution to the problem.

Then in Theorem 5, we show that the proposed algorithm satisfies the PAC constraints—i.e., it is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct for any $\epsilon_\mu, \epsilon_\rho, \delta \in (0, 1)$—, and derive an upper-bound on its sample complexity. For σ-subgaussian arms, this bound is expressed in terms of the constants

$$\begin{aligned}
C_k \triangleq \min\Bigg[&\left(\tfrac{2\sigma}{\Delta_k^*}\right)^2, \left(\tfrac{2}{\max\{0, \Delta_{\beta,k}\}}\right)^2, \\
&\max\left\{\left(\tfrac{2\sigma}{\epsilon_\mu}\right)^2, \left(\tfrac{2}{\max\{0, \epsilon_\rho - \Delta_{\beta,k}\}}\right)^2\right\}\Bigg], \qquad (25)
\end{aligned}$$

$k \in \mathcal{K}$, which characterize the complexity of concluding one of the followings: **1)** the arm violates the risk constraint or its mean is sub-optimal *(bad arm)*, or **2)** the arm nearly satisfies both conditions *(good arm)*.

Our upper-bound nearly matches (up to logarithmic factors) our general lower-bound reported in Section 3.1, and is reminiscent (up to the risk-related parameters) to the results in unconstrained PAC bandits (Even-Dar *et al.* 2006), with more finely tuned complexity term that matches our risk constraint. A slight difference in the nature of our upper and lower bounds is that in the upper-bound, μ_k is compared to $\max_{k' \in \mathcal{K}_{\beta,0}} \mu_{k'}$, whereas in the lower-bound, it is compared to the potentially larger quantity $\max_{k' \in \mathcal{K}_{\beta,\epsilon_\rho}} \mu_{k'}$. This is due to the nature of relaxation, which is essential to keep the requirements at a realistic level, and is thus unavoidable. However, this difference vanishes as ϵ_ρ approaches 0.

Before we describe and analyze the algorithm, we need to introduce a number of notations. Let $\mathcal{N}_{k,t}$ denote the set of samples drawn from arm k in the first t rounds and $N_{k,t} = |\mathcal{N}_{k,t}|$ denote its cardinality. Let $\hat{\mu}_{k,t} = \frac{1}{N_{k,t}} \sum_{X \in \mathcal{N}_{k,t}} X$, $\hat{F}_{k,t}(x) = \frac{1}{N_{k,t}} \sum_{X \in \mathcal{N}_{k,t}} \mathbf{1}(X \leq x)$, and $\hat{\rho}_{k,t}(\tau) = \inf\left\{x \in \mathbb{R} : 1 - \tau \leq \hat{F}_{k,t}(x)\right\}$ denote the empirical mean, CDF, and quantile function of arm k at round t, where $\mathbf{1}(\cdot)$ is the indicator function. The accuracy of these estimates are quantified using confidence terms $f_\mu(N) = \sqrt{\frac{2\sigma^2}{N} \ln\left(\frac{6HK}{\delta}\right)}$ and $f_\rho(N) = \sqrt{\frac{2}{N} \ln \frac{6HK}{\delta}}$.

Algorithm 1 RiskAverse-UCB-m-best

Input: quantile $\alpha \in (0,1)$, risk threshold $\beta \in \mathbb{R}$, accuracy parameters $\epsilon_\mu, \epsilon_\rho \in (0,1)$ and confidence level $\delta \in (0,1)$
for each $k \in \mathcal{K}$ **do**
  sample arm k once, and update sets $\mathcal{N}_k$, counters N_k, and estimates $\hat{\rho}_k$ and $\hat{\mu}_k$ accordingly
Set $t = |\mathcal{K}| = K$ and H as in Theorem 5
repeat
  Set $\hat{\mathcal{K}}_t = \big\{k : \hat{\rho}_{k,t}\big(\alpha - f_\rho(N_{k,t})\big) \geq \beta\big\}$
  Select an optimistic arm

$$k^\dagger_{t+1} \in \arg\max_{k \in \hat{\mathcal{K}}_t} \big(\hat{\mu}_{k,t} + f_\mu(N_{k,t})\big)$$

  Draw a sample from the selected arm $k^\dagger_{t+1}$
  Set $t = t+1$
  Update sets $\mathcal{N}_{k,t}$, counters $N_{k,t}$, and estimates $\hat{\rho}_{k,t}$ and $\hat{\mu}_{k,t}$, accordingly
until $\Big(\Big[f_\mu(N_{k^\dagger_t,t}) \leq \epsilon_\mu/2$ and $\hat{\rho}_{k^\dagger_t,t}\big(\alpha - \epsilon_\rho + f_\rho(N_{k^\dagger_t,t})\big) \geq \beta\Big]$ or $t \geq H\Big)$
return $k^\dagger_t$

Theorem 5. *Let $\mathcal{K}$ be a collection of K σ-sub-Gaussian arms for a $\sigma > 0$; $\epsilon_\mu, \epsilon_\rho \in (0,1)$ be the accuracy parameters; $\delta \in (0,1)$ be the confidence level; $\alpha \in (0,1)$ be the quantile; and $\beta \in \mathbb{R}$ be the risk threshold. We define $H \triangleq 3K\left(\frac{2\sigma^2}{\epsilon_\mu^2} + \frac{4}{\epsilon_\rho^2}\right)\ln\left[\frac{6K}{\delta}K\left(\frac{2\sigma^2}{\epsilon_\mu^2} + \frac{4}{\epsilon_\rho^2}\right)\right]$. Then, with probability at least $(1-\delta)$, Algorithm 1 outputs an arm $k^\dagger \in \mathcal{K}$ with $\mu_{k^\dagger} \geq \mu^* - \epsilon_\rho$ and $\rho_{k^\dagger}(\alpha - \epsilon_\rho) \geq \beta$ after drawing at most $\big[C_k \ln\big(\frac{6HK}{\delta}\big)\big]$ samples from each arm $k \in \mathcal{K}$, where C_k is defined as in* (25)*. In particular, Algorithm 1 is $(\epsilon_\mu, \epsilon_\rho, \delta)$-correct with sample complexity at most $\sum_{k\in\mathcal{K}} C_k \ln\big[\frac{6KH}{\delta}\big]$.*

Proof. We start the proof by defining events

$$\mathcal{E}_{\rho,1} = \Big\{\forall 1 \leq t \leq H,\ \forall k \in \mathcal{K},\ \hat{\rho}_{k,t}\big(\alpha - f_\rho(N_{k,t})\big) \leq \rho_k\big(\alpha - 2f_\rho(N_{k,t})\big)\rho(N_{k,t})\Big\},$$

$$\mathcal{E}_{\rho,2} = \Big\{\forall 1 \leq t \leq H,\ \forall k \in \mathcal{K},\ \hat{\rho}_{k,t}\big(\alpha - \epsilon_\rho + f_\rho(N_{k,t})\big) \geq \rho_k\big(\alpha - \epsilon_\rho + 2f_\rho(N_{k,t})\big)\Big\},$$

$$\mathcal{E}_{\rho,3} = \Big\{\forall 1 \leq t \leq H,\ \forall k \in \mathcal{K},\ \hat{\rho}_{k,t}\big(\alpha - f_\rho(N_{k,t})\big) \geq \rho_k(\alpha)\Big\}$$

and

$$\mathcal{E}_{\rho,4} = \Big\{\forall 1 \leq t \leq H,\ \forall k \in \mathcal{K},\ \hat{\rho}_{k,t}\big(\alpha - \epsilon + f_\rho(N_{k,t})\big) \leq \rho_k(\alpha - \epsilon)\Big\},$$

and events $\mathcal{E}_\rho = \bigcup_{i=1}^4 \mathcal{E}_{\rho,i}$ and $\mathcal{E}_\mu = \{\forall 1 \leq t \leq H,\ \forall k \in \mathcal{K},\ |\hat{\mu}_{k,t} - \mu_k| \leq f_\mu(N_{k,t})\}$. Note that $\mathcal{E}_\mu$ is the event that all the mean estimates are within the confidence interval in the first H rounds and $\mathcal{E}_\rho$ is the same for the α-quantile estimates. Using the Chernoff bound for sub-Gaussian random variables and Lemma 6 (reported at the end of this section), we can show that $\mathbb{P}[\mathcal{E}_\mu \cap \mathcal{E}_\rho] \geq 1-\delta$. From now on in this proof, we condition on the event $\mathcal{E}_\mu \cap \mathcal{E}_\rho$.

Let us define $k^* = \mathrm{argmax}_{k\in\mathcal{K}_{\beta,0}} \mu_k$. In the case that the algorithm terminates before round H, the correctness follows, because

(i) On the event $\mathcal{E}_{\rho,3}$, $k \in \hat{\mathcal{K}}_t$ for all arm k with $\rho_k(\alpha) \geq \beta$, and in particular, $k^* \in \hat{\mathcal{K}}_t$, for all $1 \leq t \leq H$.

(ii) On the event $\mathcal{E}_\mu$, the algorithm only terminates before round H, if the optimistic arm $\hat{\rho}_{k^\dagger_t,t}$ satisfies $\mu_{k^\dagger_t} + \epsilon \geq \mu_{k^\dagger_t} + 2f_\mu(N_{k^\dagger_t,t}) \geq \hat{\mu}_{k^\dagger_t,t} + f_\mu(N_{k^\dagger_t,t})$, that according to the selection rule, also upper bounds $\max_{k\in\mathcal{K}_t} \hat{\mu}_{k,t} + f_\mu(N_{k,t}) \geq \max_{k\in\mathcal{K}_t} \mu_k$, and thus, we have $\mu_{k^\dagger_t} + \epsilon \geq \mu^*$, since $k^* \in \hat{\mathcal{K}}_t$.

(iii) The algorithm only terminates before round H if $\beta \leq \hat{\rho}_{k^\dagger_t,t}\big(\alpha - \epsilon_\rho + f_\rho(N_{k^\dagger_t,t})\big)$, which on the event $\mathcal{E}_{\rho,4}$ implies that $k^\dagger_t$ satisfies the relaxed risk condition $\rho_{k^\dagger_t}(\alpha - \epsilon_\rho) \geq \beta$.

To show that the claimed sample complexity also holds on the event $\mathcal{E}_\mu \cap \mathcal{E}_\rho$, we first analyze several not necessarily disjoint cases, and then combine them to obtain the final result.

Case 1) For any arm with $\rho_k(\alpha) < \beta$, it holds that

$$\forall 1 \leq t \leq H, \quad f_\rho(N_{k,t} - 1) \geq \Delta_{k,\beta}/2\,, \tag{26}$$

which means if arm k is sampled N times, for N big enough to satisfy $f_\rho(N) < \Delta_{k,\beta}/2$, then its quantile estimate clearly shows that it violates the risk constraint and gets discarded. In fact, assume that $f_\rho(N_{k,t}) < \Delta_{k,\beta}/2$ holds for some $1 \leq t \leq H$ (note that $\Delta_{k,\beta} \geq 0$ when $\rho_k(\alpha) < \beta$). Then, on the event $\mathcal{E}_{\rho,1}$, we have $\hat{\rho}_{k,t}\big(\alpha - f_\rho(N_{k,t})\big) \leq \rho_k\big(\alpha - 2f_\rho(N_{k,t})\big) \leq \rho_k(\alpha - \Delta_{k,\beta}) = \inf\big\{x \in \mathbb{R} : 1 - \alpha + \Delta_{k,\beta} \leq F_k(x)\big\} = \inf\big\{x \in \mathbb{R} : F_k(\beta) \leq F_k(x)\big\} \leq \beta$, and thus, arm k is discarded from $\hat{\mathcal{K}}_t$ for the rest of the rounds.

Case 2) For any arm with $\beta \leq \rho_k(\alpha - \epsilon_\rho)$, it holds that

$$\forall 1 \leq t \leq H,\ \Big(f_\rho(N_{k,t}) \geq \epsilon_\rho/2 - \Delta_{k,\beta}/2 \quad \text{or} \quad \hat{\rho}_{k,t}(\alpha - \epsilon_\rho + f_\rho(N_{k,t})) \geq \beta\Big)\,,$$

which means if arm k is sampled N times for N big enough to satisfy $f_\rho(N) < \epsilon_\rho/2 - \Delta_{k,\beta}/2$, then its quantile estimate clearly shows that arm k satisfies the relaxed risk constraint. To see why this is true, note that $f_\rho(N_{k,t}) < \epsilon_\rho/2 - \Delta_{k,\beta}/2$ implies $1 - \alpha + (\epsilon_\rho - 2f_\rho(N_{k,t})) > (1-\alpha) + \Delta_{k,\beta} = F_k(\beta)$, and thus, on the event $\mathcal{E}_{\rho,2}$, we have $\hat{\rho}_{k,t}\big(\alpha - \epsilon_\rho + f_\rho(N_{k,t})\big) \geq \rho_k\big(\alpha - \epsilon_\rho + 2f_\rho(N_{k,t})\big) = \inf\big\{x : 1 - \big(\alpha - \epsilon_\rho + 2f_\rho(N_{k,t})\big) \leq F_k(x)\big\} > \beta$ (note that in this case $\Delta_{\beta,k} \leq \epsilon_\rho$).

Case 3) For any arm k with $\mu_k \leq \mu^*$, it holds that

$$\forall 1 \leq t \leq H, \quad f_\mu(N_{k,t} - 1) > (\mu^* - \mu_k)/2\,, \tag{27}$$

which means if arm k is sampled large enough time to have $f_\mu(N_{k,t}) < (1/2)(\mu^* - \mu_k)$, it is concluded that its mean is not big enough, and the algorithm does not sample it anymore. In fact, assume that $f_\mu(N_{k,t}) < (1/2)(\mu^* - \mu_k)$ for some $1 \leq t \leq H$. Then, on event $\mathcal{E}_\mu$, we have $\hat{\mu}_{k,t} + f_\mu(N_{k,t}) \leq \mu_k + 2f_\mu(N_{k,t}) < \mu^*$. Therefore, as $\hat{\mu}_{k^*,t'} + f_\mu(N_{k^*,t'}) \geq \mu^*$ for each $1 \leq t' \leq H$, according to the selection rule, arm k won't be sampled for the rest of the rounds.

Case 4) For any arm k with $\mu_k > \mu^*$ and $\forall 1 \leq t \leq H,\ f_\mu(N_{k,t}) > \Delta_k^*$, we have that $\Delta_k^* = 0$.

Case 5) For any arm k with $\rho_k(\alpha) \geq \beta$ and $\forall 1 \leq t \leq H,\ f_\rho(N_{k,t}) > \Delta_{\beta,k}$, we have that $\Delta_{\beta,k} \leq 0$.

Case 6) For any arm k with $\mu_k > \mu^* - \epsilon_\mu$, if at round t the selected arm $k = k_t^\dagger$ satisfies $2f_\mu(N_{k,t}) < \epsilon_\rho$, then $\hat{\rho}_{k,t}(\alpha - \epsilon_\rho + f_\rho(N_{k,t})) < \beta$ or the algorithm terminates immediately and returns arm k.

We conclude the proof by investigating the following cases:

Case (a) An arm k with both $\mu^* - \epsilon \leq \mu_k$ and $\beta \leq \rho(\alpha - \epsilon_\rho)$ is not sampled more than

$$\min\left[\left(\tfrac{2\sigma}{\Delta_k^*}\right)^2, \left(\tfrac{2}{\max\{0,\Delta_{\beta,k}\}}\right)^2, \max\left\{\left(\tfrac{2\sigma}{\epsilon_\mu}\right)^2, \left(\tfrac{2}{\epsilon_\rho - \Delta_{\beta,k}}\right)^2\right\}\right] \ln\left(\tfrac{6HK}{\delta}\right)$$

times. Note also that in this case $0 \leq \Delta_{\beta,k} \leq \epsilon_\rho$, and thus, $\epsilon_\rho - \Delta_{\beta,k} = \max\{0, \epsilon_\rho - \Delta_{\beta,k}\}$, which implies that $C_k \ln(6KH/\delta)$ is indeed an upper-bound, and also that $C_k \leq 2\sigma/\epsilon_\mu^2 + 4/\epsilon_\rho^2$.

Case (b) An arm k with $\rho_k(\alpha - \epsilon_\rho) < \beta$ cannot be the output if the algorithm terminates before round H. Besides, it is sampled at most $\min\left\{\left(\tfrac{2\sigma}{\Delta_k^*}\right)^2, \left(\tfrac{2}{\Delta_{\beta,k}}\right)^2\right\} \ln\left(\tfrac{6HK}{\delta}\right)$ times. However, for all these arms, it holds that $\Delta_{\beta,k} \geq \epsilon_\rho$, and thus, it also holds that $\max\left\{\left(\tfrac{2\sigma}{\epsilon_\mu}\right)^2, \left(\tfrac{2}{\max\{0,\epsilon_\rho - \Delta_{\beta,k}\}}\right)^2\right\} = \infty$. As a result, $C_k \ln(6KH/\delta)$ is indeed an upper-bound. It also follows that $C_k \leq 2\sigma/\epsilon_\mu^2 + 4/\epsilon_\rho^2$.

Case (c) An arm k with $\mu_k < \mu^* - \epsilon_\mu$ cannot be the output if the algorithm terminates before round H. Besides, it is sampled at most $\min\left\{\left(\tfrac{2\sigma}{\Delta_k^*}\right)^2, \left(\tfrac{2}{\max\{0,\Delta_{\beta,k}\}}\right)^2\right\} \ln\left(\tfrac{6HK}{\delta}\right)$ times. Note that $\mu_k < \mu^* - \epsilon_\mu$ implies $\Delta_k^* \geq \epsilon_\mu$, and thus, $\left(\tfrac{2\sigma}{\Delta_k^*}\right)^2 \leq \max\left\{\left(\tfrac{2\sigma}{\epsilon_\mu}\right)^2, \left(\tfrac{2}{\max\{0,\epsilon_\rho - \Delta_{\beta,k}\}}\right)^2\right\}$, implies that $C_k \ln(6KH/\delta)$ is indeed an upper-bound. It also follows that $C_k \leq 2\sigma/\epsilon_\mu^2 + 4/\epsilon_\rho^2$.

All that left is to show that the algorithm terminates before round H. This follows from the fact that C_k is upper-bounded by $\max\{\epsilon_\mu^{-1}, \epsilon_\rho^{-1}\}$ due to the claims given in the three cases **(a)-(c)**, and the fact that $H \geq K \max\{\epsilon_\mu^{-1}, \epsilon_\rho^{-1}\} \ln(\frac{6HK}{\delta})$ because $3c\ln(cr) = c\ln(c^3 r^3) \geq c\ln(r[3c\ln(cr)])$, for any $c, r > 0$ such that $cr \geq 6$. □

Lemma 6. *For any $t \geq 1$, $k \in \mathcal{K}$, and $\tau \in (0,1)$, it holds that $\mathbb{P}\big(\rho_k(\tau) < \hat{\rho}_{k,t}(\tau + \Delta)\big) \leq \exp(\Delta^2 N_{k,t}/2)$, for $\Delta \in (0, 1-\tau)$, and $\mathbb{P}\big(\rho_k(\tau) > \hat{\rho}_{k,t}(\tau - \Delta)\big) \leq \exp(\Delta^2 N_{k,t}/2)$, for $\Delta \in (0, \tau)$.*

The proof is based on the Chernoff bound and is omitted due to space constraints,

5 Discussion and Concluding Remarks

As discussed in the paper, our lower-bound for the case of Gaussian arms is stronger than that for the general case in the sense that it applies to any instance of the Gaussian problem, whereas the general lower-bound only applies to special problem instances. Although it would be nice to have a lower-bound in the general case that applies to all problem instances, it is not even clear what kind of interdependency of the distributions we should expect to appear in such a bound. This is in contrast to the Gaussian case, where the nice underlying structure provides easy to exploit relations.

Regarding the tightness of our results, it should be noted that the classic lower and upper bounds for risk-neutral bandits that our results are based on (Even-Dar *et al.* 2006; Mannor and Tsitsiklis 2004), are known to be suboptimal. In particular, it was shown for a median-elimination-based method that the $\log$ factor in the sample complexity upper bound can be replaced by a $\log\log$ factor (Karnin *et al.* 2013). Later, a matching lower bound was presented for a special case (Jamieson *et al.* 2014). The same paper also proposed a novel, index-based algorithm with much tighter confidence bound sequences. This was also shown to result in an improved empirical performance. Subsequently, the distance-based Δ terms in the sample complexity upper bound were also shown to be replaceable by the so called Chernoff-information, as demonstrated by another index-based method (Tánczos *et al.* 2017). In order to obtain similar results in the risk-averse setting, one should derive first similar confidence bounds for the risk measure. However, it should also be noted that the above mentioned index-based methods perform best arm identification only, and have no adaptation for the relaxed problem yet that can halt early when a *near*-optimal solution is found.

Finally, an important novelty that the risk-averse setup offers is that it makes sense to work with linear (more precisely, convex) combinations of the arms. In the classical bandit setting such a relaxation did not provide any benefit; however, in our risk-constrained formulation, it is quite possible that an arm that is too risky by itself, can be an integral part of the optimal solution, if its mean reward is high enough. This is related to the Markowitz model in the case of mean-variance optimization.

6 Acknowledgement

This research was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL). A portion of this work was completed when Balázs Szörényi was a postdoc at Technion, Israel.

References

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Jean-Yves Audibert, Sébastien Bubeck, and Remi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 41–53, 2010.

Patrick Billingsley. Probability and measure. *A Wiley-Interscience Publication, John Wiley & Sons, New York*, 1995.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25*, pages 3212–3220, 2012.

Nicolas Galichet, Michèle Sebag, and Olivier Teytaud. Exploration vs. exploitation vs. safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

W. Haskell, H. Xu, Q. Chao, and Y. Zhiyue. Online risk-aware optimization. 2016.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 423–439, Barcelona, Spain, 13–15 Jun 2014. PMLR.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pages 655–662, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1238–1246, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 218–233, 2013.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

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

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3284–3292, 2012.

Ervin Tánczos, Robert Nowak, and Bob Mankoff. A kl-lucb algorithm for large-scale crowdsourcing. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5896–5905. Curran Associates, Inc., 2017.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

William R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935.

S. Vakili and Q. Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, Sept 2016.

Jia Yuan Yu and Evdokia Nikolova. Sample complexity of risk-averse bandit-arm selection. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2576–2582, 2013.

Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *CoRR*, abs/1405.0833, 2014.