Index-Based Policy for Risk-Averse Multi-Armed Bandit

September 17, 2018

Authors

Jianyu Xu: Department of Industrial Systems Engineering and Management, National University of Singapore, 1 Engineering Drive 2, Singapore.
William B. Haskell: Department of Industrial Systems Engineering and Management, National University of Singapore, 1 Engineering Drive 2, Singapore.
Zhisheng Ye: Department of Industrial Systems Engineering and Management, National University of Singapore, 1 Engineering Drive 2, Singapore.

Corresponding author

William B. Haskell: wbhaskell@gmail.com.
Abstract

The multi-armed bandit (MAB) is a classical online optimization model for the trade-off between exploration and exploitation. The traditional MAB is concerned with finding the arm that minimizes the mean cost. However, minimizing the mean does not take the risk of the problem into account. We now want to accommodate risk-averse decision makers. In this work, we introduce a coherent risk measure as the criterion to form a risk-averse MAB. In particular, we derive an index-based online sampling framework for the risk-averse MAB. We develop this framework in detail for three specific risk measures, i.e. the conditional value-at-risk, the mean-deviation and the shortfall risk measures. Under each risk measure, the convergence rate for the upper bound on the pseudo regret, defined as the difference between the expectation of the empirical risk based on the observation sequence and the true risk of the optimal arm, is established.

Keywords: Stochastic programming, multi-armed bandit, online optimization, coherent risk, index policy.

1 Introduction

The multi-armed bandit (MAB) is a classical problem named after an imaginary slot machine with multiple arms (Robbins 1985). At each time step, a player pulls one of the arms and receives a random cost associated with the chosen arm. A policy, or a strategy for the player, denoted by $\varphi$, is an algorithm to choose the next arm based on the past choices and the observed cost sequence. Given a sequence of pulls from the arms, the pseudo regret, which measures the performance of the policy, is defined as the difference between the expectation of the average cost after $n$ time steps and the lowest mean cost among the arms. The player’s objective is to design a policy that minimizes the pseudo regret either asymptotically (Lai and Robbins 1985) or uniformly over time (Auer et al. 2002). The above basic MAB has many variations depending on the application. Recent work on this topic has led to many important theoretical results (Agrawal et al. 1988; Bubeck et al. 2012) as well as interesting applications (Jennison and Turnbull 1999; Mohri and Munoz 2014; Scott 2015). The asymptotically optimal policy is usually obtained by the index-based policy proposed in Lai and Robbins (1985), and extended by Agrawal (1995), Auer et al. (2002) and Kleinberg (2005). An index-based policy calculates an index for each arm at each time step. An index consists of two parts, the empirical estimation of the mean and another term related to the probability confidence bound from the estimation. The arm with the lowest current index is chosen each time.

The classical MAB defines the pseudo regret in terms of the long-term average cost, and thus it is risk-neutral. The risk-neutral MAB does not take the variance of the random cost of each arm into consideration. As a result, it is inappropriate in many applications which require reliability guarantees. In clinical trials, for example, a specific treatment with a low average side effect can be highly variable among different patients and thus it may cause serious medical problems for an individual. These applications require a criterion that takes the variation/spread/ dispersion of the arms into account. It is thus natural to select a risk measure and then to try to choose the arm with the lowest risk. Some popular risk measures include value-at-risk (Benati and Rizzi 2007), conditional value-at-risk (CVaR) (Rockafellar et al. 2000)
and expected shortfall (Acerbi and Tasche 2002). Risk measures have been extensively studied in decision analysis (Föllmer and Schied 2002; Brown and Sim 2009; Brown et al. 2012), reinforcement learning (Mannor and Tsitsiklis 2011; Shen et al. 2013) and operations research (Ruszczyński and Shapiro 2006; Liu et al. 2017).

The risk-averse MAB has attracted significant attention in the recent literature. Sani et al. (2012) uses the mean-variance risk measure to assess the performance of each arm. They proposed an index-based policy and proved a sub-linear upper bound for the pseudo regret. Maillard (2013) extends the mean-variance measure in Sani et al. (2012) to a more general risk measure that takes the tail of the cumulative distribution function (CDF) into consideration. Galichet et al. (2013) introduces the conditional value-at-risk to MAB. However, their theoretical analysis only focuses on the special case where the CVaR degenerates to the essential infimum. These studies adopt the index-based policy inherited from the classical risk-neutral MAB. The risk measures involved in existing work are limited and many popular risk measures are not included. So far, there is a lack of research on risk-averse MAB and there is not yet a consensus on the notion of pseudo regret. In the risk-neutral MAB, the loss of a single choice is measured as the difference between the means of the chosen arm and the optimal arm. The total regret is the summation of the losses of each single choice in a sequence. The reason is because the total cost is the summation of the cost drawn each time. So, the total regret is also additive in terms of each single loss. However, when we use a risk measure as the criterion in the risk-averse MAB, the property of additivity no longer holds. This leads to the requirement of a rational pseudo regret for the risk-averse MAB.

In this work, we formalize the risk-averse MAB by generalizing the classical risk-neutral MAB. We introduce a general coherent risk measure in the MAB as the criterion for the optimal arm. We define the pseudo regret as the expectation of the difference between the empirical risk measure based on the observation sequence and the true risk measure of the optimal arm. Further, we conclude that by using our notion of pseudo regret as the target function for optimization, we are still correctly searching for the single optimal arm. An index-based policy is then constructed to find the optimal risk-averse arm. To prove the convergence of the pseudo regret under the proposed policy, we restrict the general risk measure to three specific cases: that is, the CVaR, the mean-deviation (MD) and the shortfall risk measure. These risk measures have not yet been fully investigated for the risk-averse MAB. To show the performance of the policy in the long run, we present a detailed theoretical analysis of the convergence rate of the pseudo regret for each of these three risk measures.

This paper is organized as follows. Section 2 formalizes the risk-averse MAB by using a general coherent risk measure as the decision maker’s criterion for arm selection. We define the pseudo regret and then construct an index-based policy for a general risk-averse MAB. In Section 3, we introduce our three specific risk measures of interest. Then, we present the main results of this work; that is, the expression of the index and the corresponding convergence rate of the pseudo regret for each of these three risk measures. Section 4 provides the detailed proofs of our main results. Some concluding remarks are given in Section 5.
2 Risk-Averse MAB

Consider an MAB with $K$ arms in total. Each pull of an arm $k$, $k = 1, \ldots, K$, generates a realization of a nonnegative random cost $X_k$ with mean $\mu_k \triangleq \mathbb{E}(X_k)$ and cumulative distribution function $F_k$. In the remainder of this work, we suppose all $X_k$’s are essentially bounded by $M$, i.e. $\mathbb{P}\{X_k \in [0, M]\} = 1$ for all $k = 1, \ldots, K$. This assumption is commonly used in the MAB literature, see Sani et al. (2012) and Maillard (2013) among others. We further suppose that successive pulls from an arm yield a sequence of i.i.d. random costs and each pull does not change the distributions of the $K$ arms.

We operate on a finite time horizon of $n$ time steps. A policy $\varphi$ generates a sequence of choices $\{I_t\}_{t=1}^n$ where at each time $t \in \{1, \ldots, n\}$, arm $I_t \in \{1, \ldots, K\}$ is chosen. Let the number of observations from arm $k$ up to time $n$ be $T_k(n)$, i.e. $T_k(n) \triangleq \sum_{t=1}^n I\{I_t = k\}$, where $I\{\cdot\}$ is the indicator function. The risk-neutral MAB problem seeks the arm $k^*$ with the lowest mean cost, i.e. $k^* \triangleq \arg\min_{k=1, \ldots, K} \mu_k$. Equivalently, it attempts to minimize the following pseudo regret (Lai and Robbins 1985)

$$
\max_{k=1, \ldots, K} \left\{ \mathbb{E}\left(\frac{1}{n} \sum_{t=1}^n x_{I_t, t} - \mu_k \right) \right\} = \frac{1}{n} \sum_{k=1}^K \mathbb{E}T_k(n) (\mu_k - \mu_{k^*}).
$$

(1)

2.1 Risk-averse MAB formulation

To formalize the risk-averse MAB, we introduce a coherent risk-measure $\rho$ as the decision maker’s objective instead of the expectation. A coherent risk measure is defined as follows.

**Definition** (Ruszczyński and Shapiro 2006) Let $\mathcal{L}$ be a space of essentially bounded random variables. A risk measure $\rho : \mathcal{L} \to \mathbb{R}$ is called a coherent risk measure if it satisfies:

1. (1.1) Convexity: For all $X, Y \in \mathcal{L}$ and $\forall \lambda \in [0, 1]$, $\rho [\lambda X + (1 - \lambda) Y] \leq \lambda \rho (X) + (1 - \lambda) \rho (Y)$;

2. (1.2) Monotonicity: If $X, Y \in \mathcal{L}$ and $X \leq Y$, then $\rho (X) \leq \rho (Y)$;

3. (1.3) Translation equivalence: If $\alpha \in \mathbb{R}$ and $X \in \mathcal{L}$, then $\rho (X + \alpha) = \rho (X) + \alpha$;

4. (1.4) Positive homogeneity: If $\beta > 0$ and $X \in \mathcal{L}$, then $\rho (\beta X) = \beta \rho (X)$.

Let $\rho_k \triangleq \rho(X_k)$ denote the risk of arm $k$. In the risk-averse MAB, we define the optimal arm to be the one with the lowest risk, i.e. $k^* \triangleq \arg\min_{k=1, \ldots, K} \rho_k$. We assume that the optimal arm $k^*$ is always unique. We call arm $k$ a sub-optimal arm if $k \neq k^*$. This assumption is reasonable in practice, as it is uncommon for risk measures for two different populations to coincide. Three specific risk measures will be introduced and studied in Section 3.

Our first objective is to extend the definition of the pseudo regret from the risk-neutral MAB to the risk-averse MAB. From (1), we note that the risk-neutral MAB aims to minimize the expectation of the difference between the empirical mean of the whole observation sequence and the mean of the optimal arm. Similarly, in the risk-averse case, we can let $\hat{\rho}_{\varphi, n}$ be the empirical risk based on $\{x_{I_t, t}\}_{t=1}^n$. Then, we can use the expectation of the difference between $\hat{\rho}_{\varphi, n}$ and $\rho_{k^*}$ as the pseudo regret. A formal definition of the pseudo regret is given below.
**Definition** The pseudo regret of any given policy \( \varphi \) under a risk measure \( \rho \) is defined as

\[
R_n(\varphi) \triangleq E[\hat{\rho}_{\varphi, n} - \rho_k^*].
\]

(2)

Minimizing \( R_n(\varphi) \) yields the arm with the lowest risk asymptotically. Throughout this work, we call \( \hat{\rho}_{\varphi, n} \) the empirical risk measure of the policy \( \varphi \).

2.2 Algorithm for the risk-averse MAB

In this section, we present our main algorithm for the risk-averse MAB. The main idea of the algorithm is based on the notion of the lower probability confidence bound. Therefore, we call it the Risk-Averse Lower Confidence Bound (RA-LCB) algorithm. To initialize, the algorithm chooses each arm once. This ensures that each term \( T_k(n) \) is larger than 0 after initialization. The calculation of the index for each arm after initialization relies on the condition \( T_k(n) > 0 \).

At each time \( n > K \), the algorithm calculates an index for each arm \( k, k = 1, \ldots, K \), which is the difference between two terms \( \hat{\rho}_{k, n} \) and \( \varepsilon_\rho(\cdot) \). The first term is the empirical estimate of \( \rho_k \). The second term \( \varepsilon_\rho(\cdot) \) is a function of \( n, T_k(n), K \) and the confidence level \( \delta \in (0, 1) \). It is related to the probability confidence bound of the empirical estimate \( \hat{\rho}_{k, n} \). This second term ensures that the true risk of an arm falls above its index with an overwhelming probability as \( n \) becomes larger. This is the basis of the convergence proof for the pseudo regret. Let \( S_k(n) \triangleq \{ t : I_t = k, t \leq n \} \) be the set of all times at which arm \( k \) is chosen. Then \( X_{k,t}, \; t \in S_k(n) \) is the observation sequence from arm \( k \). Both terms in the index are functions of the sequence \( X_{k,t}, \; t \in S_k(n) \). The specific forms of these two terms vary under different risk measures. We will specify both terms under the three risk measures that we consider in the sections below. As in the risk-neutral case, at each time step, the arm with the lowest index is chosen. A summary of the main flow of our procedure is given in Algorithm 1. Throughout the rest of the paper, we fix the confidence level \( \delta \in (0, 1) \).

**Algorithm 1** Risk-Averse Lower Confidence Bounds (RA-LCB)

**Input:** constant \( \delta, K \), a confidence bound function \( \varepsilon_\rho(\cdot) \).

**Initialization:** Choose each arm once in the first \( K \) pulls.

while \( n > K \)

for \( k = 1 : K \) do

1. Calculate \( \hat{\rho}_{k, n} \) based on \( X_{k,t}, \; t \in S_k(n) \).

2. Calculate the index for each arm: \( B_{k, n} \triangleq \hat{\rho}_{k, n} - \varepsilon_\rho(n, T_k(n), K, \delta) \).

end for

\( I_n = \arg\min_k B_{k, n} \).

\( n \leftarrow n + 1 \).

pull arm \( I_n \) at time \( n, n = 1, 2, \ldots \).

end while
3 Main results

This section presents our main results. We apply Algorithm 1 to three specific risk measures: CVaR, MD and shortfall. For each risk measure, we derive the corresponding upper bound for its pseudo regret. We need the following condition when we establish the convergence rate of the pseudo regret for the three risk measures. The explicit form of \( \varepsilon_{\rho}(\cdot) \) under each risk measure is given below. In Section 4, we will show that each specific \( \varepsilon_{\rho}(\cdot) \) we give in this section satisfies this condition though the procedure of proof.

**Condition 1** For all \( k = 1, \ldots, K \), the function \( \varepsilon_{\rho}(\cdot) \) satisfies

\[
P \{ |\hat{\rho}_{k,n} - \rho_k| \geq \varepsilon_{\rho}(n, T_k(n), K, \delta) \} \leq \frac{C \delta}{n^2 K},
\]

where \( C \) is a constant.

The construction of \( \varepsilon_{\rho}(\cdot) \) under each specific risk measure is based on the various forms of concentration results for \( |\hat{\rho}_{k,n} - \rho_k| \). Specifically, \( \varepsilon_{\rho}(\cdot) \) is directly related to the confidence bounds derived from certain concentration inequalities. If the empirical risk \( \hat{\rho}_{k,n} \) is a summation of different terms, we derive the confidence bound on each term and use the summation of these bounds to achieve \( \varepsilon_{\rho}(\cdot) \). Under Condition 1, the convergence rate of the pseudo regret for each risk measure is of the order \( O(\sqrt{\log n/n}) \). The upper bound for the pseudo regret under each risk measure is correspondingly given in Section 3.1-3.3. In Section 3.4, we make some discussion on the order \( O(\sqrt{\log n/n}) \) of the convergence rate.

3.1 Conditional value-at-risk

We first define CVaR as follows.

**Definition (CVaR risk measure)** Consider a random variable \( Y \) with CDF \( F_Y \). For a fixed level \( \alpha \in [0, 1] \), let \( F_Y^{-1}(\alpha) \triangleq \inf \{ x : F_Y(x) \geq \alpha \} \) be the \( \alpha \)-quantile of \( Y \). Then, the CVaR at level \( \alpha \) of arm \( k \) is \( \mathbb{E}[X_k | X_k \geq F_Y^{-1}(\alpha)] \), i.e.

\[
\rho^C_k \triangleq (1 - \alpha)^{-1} \int_\alpha^1 F_Y^{-1}(\tau)d\tau.
\]

When \( \alpha = 0 \), CVaR becomes expectation. We need the following assumption for CVaR which ensures the convergence of the empirical estimate of the risk towards its actual value. Lemma 1 below gives a convenient representation of CVaR and also relies on this assumption.

**Assumption 1.** For all \( k = 1, \ldots, K \), \( F_k \) is continuously differentiable on \((0, M)\) with corresponding density function \( f_k \) and \( |f_k(F_k^{-1}(\alpha))|^{-1}, k = 1, \ldots, K \) is uniformly upper bounded by a constant \( m(\alpha), \forall \alpha \in (0, 1) \).

This common assumption also appears in some statistical literature (Bahadur 1966; Arcones et al. 1996). In this work, we use a convenient representation of CVaR (Rockafellar et al. 2000) given as

\[
\rho^C_k = \inf_{\eta \in \mathbb{R}} \{ \eta + (1 - \alpha)^{-1} \mathbb{E}[(X_k - \eta)_+] \}.
\]
As \( \eta + (1 - \alpha)^{-1} \mathbb{E}[(X_k - \eta)_+] \) is convex in \( \eta \), the range of \( \eta \) can be restricted to the support of \( F_k \), as shown in the following lemma.

**Lemma 1.** ([Rockafellar et al. 2000] Theorem 1) Consider a random variable \( X \) with a continuous CDF \( F_X \) and bounded support \([0, M]\). Let

\[
g(\eta) \triangleq \eta + (1 - \alpha)^{-1} \mathbb{E}[(X - \eta)_+],
\]

then \( g(\eta) \) reaches its global minimal at \( \eta = F_X^{-1}(\alpha) \). Further, the CVaR of \( X \) is

\[
F_X^{-1}(\alpha) + (1 - \alpha)^{-1} \mathbb{E}[(X - F_X^{-1}(\alpha))_+].
\]

From the sequence \( X_{k,t}, t \in S_k(n) \) of arm \( k \) as defined in Algorithm 1, the empirical distribution for \( X_k \) up to time \( n \) is

\[
F_{k,T_k(n)}(x) \triangleq \frac{1}{|S_k(n)|} \sum_{t \in S_k(n)} I\{x_{k,t} \leq x\},
\]

where \( | \cdot | \) denotes the number of items in a set. Similarly, from the sequence \( \{x_{I_t,t}\}_{t=1}^n \) drawn by the policy \( \varphi \), we define

\[
F_{\varphi,n}(x) \triangleq \frac{1}{n} \sum_{t=1}^n I\{x_{I_t,t} \leq x\}.
\]

The empirical estimate \( \hat{\rho}_{k,n} \) of the CVaR of arm \( k \) may then be defined as

\[
\hat{\rho}_{k,n}^C \triangleq F_{k,T_k(n)}^{-1}(\alpha) + (1 - \alpha)^{-1} \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} [x_{k,t} - F_{k,T_k(n)}^{-1}(\alpha)]_+,
\]

and the empirical estimate \( \hat{\rho}_{\varphi,n} \) of the CVaR for the policy \( \varphi \) may be defined as

\[
\hat{\rho}_{\varphi,n}^C \triangleq F_{\varphi,n}^{-1}(\alpha) + (1 - \alpha)^{-1} \frac{1}{n} \sum_{t=1}^n [x_{I_t,t} - F_{\varphi,n}^{-1}(\alpha)]_+.
\]

According to (2), the pseudo regret under the CVaR up to time \( n \) is then

\[
R_{n}^C(\varphi) \triangleq \mathbb{E}\hat{\rho}_{\varphi,n}^C - \rho_{k}^C.
\]

For a fixed \( \alpha \), let \( \eta_k \triangleq F_k^{-1}(\alpha) \), \( \eta_k,n \triangleq F_{k,n}^{-1}(\alpha) \), \( \eta_{\varphi} \triangleq F_{\varphi,n}^{-1}(\alpha) \), \( \eta^* \triangleq \eta_{\varphi} \) and \( \eta^* \triangleq \eta_{\varphi,n} \). Using these notations, we can represent the pseudo regret in the following explicit form

\[
R_{n}^C(\varphi) \triangleq \mathbb{E}(\eta_{\varphi,n} - \eta^*) + (1 - \alpha)^{-1} \frac{1}{n} \sum_{k=1}^K \mathbb{E}T_k(n)\mathbb{E}[(X_k - \eta_{\varphi,n})_+ - (X_k^* - \eta^*)_+].
\]

The term \( \varepsilon_p(n, T_k(n), K, \delta) \) for CVaR case is defined as follows

\[
\varepsilon_p^C(n, T_k(n), K, \delta) \triangleq \left(1 - \alpha\right)^{-1} \left(1 - \frac{3\delta}{n}\right) M + 2 \left[1 + (1 - \alpha)^{-1}\right] m(\alpha) \sqrt{\frac{\log 2 n^2 K}{2T_k(n)}}
\]
The term $\varepsilon^C_\rho$ consists of two parts. They are the corresponding confidence bounds on the two terms in $\hat{\rho}^C_{k,n}$ respectively. Thus, the index for arm $k$ at time $n$ for CVaR is

$$B^C_{k,n} \triangleq \hat{\rho}^C_{k,n} - \varepsilon^C_\rho(n, T_k(n), K, \delta).$$

The upper bound for the pseudo regret under CVaR is established in the following theorem. We let

$$M^\phi(n) \triangleq \max\{\alpha, 1 - \alpha\} \, O\left(\frac{\log n}{n}\right) m(\alpha) + 2m(\alpha) \sqrt{\frac{\log \frac{4n}{2n}}{2n}},$$

and

$$M_k(n) \triangleq 2\log \frac{2n^2K}{\delta} \left(\frac{2\left[1 + (1 - \alpha)^{-1}\right] m(\alpha) + (1 - \alpha)^{-1} M}{\Delta^C_k - (1 - \alpha)^{-1} \frac{44}{n^2K} M}\right)^2 + 3\delta,$$

where $\Delta^C_k \triangleq \rho^C_k - \rho^C_{k*}$.

**Theorem 1.** For all $n > K$, the pseudo regret $R^C_n(\phi)$ satisfies

$$R^C_n(\phi) \leq (1 - \frac{4\delta}{n})[M^\phi(n) + (1 - \alpha)^{-1} \frac{1}{n} \sum_{k=1}^K M_k(n)(M + \Delta^C_k)]$$

$$+ \frac{4\delta}{n}[(1 - \alpha)^{-1} + 1]M + (1 - \alpha)^{-1} \Delta^C_k].$$

Specifically, the order of the upper bound is dominated by the term $M^\phi(n)$ which is $O(\sqrt{\log n/n})$.

### 3.2 Mean-deviation risk

We define the MD as follows.

**Definition (MD risk measure)** The mean-deviation of arm $k$ is

$$\rho^M_k \triangleq \mu_k + \gamma \|X_k\|_p,$$

where $\|X_k\|_p \triangleq (E|X_k - \mu_k|^p)^{\frac{1}{p}}$, $\gamma \geq 0$ and $p \in [1, \infty)$.

The constant $\gamma$ is a coefficient controlling the trade-off between the mean and the $L_p$ norm. When $\gamma = 0$, the MD measure becomes expectation; in the special case of $p = 2$, the second term becomes the standard deviation. Based on the observation sequence of arm $k$, the empirical estimate $\hat{\rho}^M_{k,n}$ of the MD risk of arm $k$ may be defined as

$$\hat{\rho}^M_{k,n} \triangleq \bar{x}_{k,T_k(n)} + \gamma \left(\frac{1}{|S_k(n)|} \sum_{t \in S_k(n)} |x_{k,t} - \bar{x}_{k,T_k(n)}|^p\right)^{\frac{1}{p}},$$

where $\bar{x}_{k,T_k(n)} \triangleq \frac{1}{|S_k(n)|} \sum_{t \in S_k(n)} x_{k,t}$. Based on the whole observation sequence the empirical estimate $\hat{\rho}^M_{\phi,n}$ of the MD risk for the policy $\phi$ may be defined as

$$\hat{\rho}^M_{\phi,n} \triangleq \bar{x}_{\phi,n} + \gamma \left(\frac{1}{n} \sum_{t=1}^n |x_{I_t, t} - \bar{x}_{\phi,n}|^p\right)^{\frac{1}{p}},$$
where \( \bar{x}, n = \frac{1}{n} \sum_{t=1}^{n} x_{I_t,t} \). The pseudo regret under the MD up to time \( n \) is then

\[
R_n^M(\varphi) \triangleq \left[ E(\bar{x}, n) - \mu_{k^*} \right] + \gamma \left[ E \left( \frac{1}{n} \sum_{t=1}^{n} |x_{I_t,t} - \bar{x}, n|^p \right)^{\frac{1}{p}} - (E |X_{k^*} - \mu_{k^*}|^p)^{\frac{1}{p}} \right].
\]

The term \( \varepsilon_{\rho}(n, T_k(n), K, \delta) \) for MD is defined as below

\[
\varepsilon_{\rho}^M(n, T_k(n), K, \delta) \triangleq M \sqrt{\log \frac{4n^2K}{\delta}} - M \left[ (p+1) \sqrt{\log \frac{4n^2K}{\delta}} \right]^{\frac{1}{p}}.
\]

Similar to the CVaR case, \( \varepsilon_{\rho}^M \) is also a summation of two terms. The are the corresponding bounds for the empirical mean and empirical \( L_p \) norm in \( \hat{\rho}_{k,n}^M \) respectively. So the index for arm \( k \) at time \( n \) for MD is

\[
B_{k,n}^M \triangleq \hat{\rho}_{k,n}^M - \varepsilon_{\rho}^M(n, T_k(n), K, \delta).
\]

The upper bound for the pseudo regret under the MD is given in the following theorem. We let

\[
N_{\varphi}(n) \triangleq \sum_{k \neq k^*} \left( 1 - \frac{\delta}{n} \right) m_k \log \frac{4n^2K}{\delta} + (k-1) \delta,
\]

where \( m_k \) will be specified in Lemma 9 below.

**Theorem 2.** For all \( n > K \), the pseudo regret satisfies

\[
R_n^M(\varphi) \leq \left( \Delta_k^M + 2\gamma p M^p \right) \frac{N_{\varphi}(n)}{n} + M \left( \frac{\delta}{n} + \left( 1 - \frac{\delta}{n} \right) \sqrt{\frac{\log \frac{4n^2K}{\delta}}{2(n - N_{\varphi}(n))}} \right),
\]

where \( \Delta_k^M \triangleq |\mu_k - \mu_{k^*}|. \) Specifically, the order of the upper bound is dominated by the term

\[
\sqrt{\log(4n^2K/\delta)[2(n - N_{\varphi}(n))^{-1}]},
\]

which is of the order \( O(\sqrt{\log n/n}) \).

### 3.3 Shortfall risk

We define shortfall risk as follows.

**Definition (Shortfall risk measure)** Let \( l : \mathbb{R} \to \mathbb{R} \) be a convex loss function, the shortfall risk measure of arm \( k \) is defined as

\[
\rho_k^S \triangleq \inf \{ \kappa \in \mathbb{R} : E[l(X_k - \kappa)] \leq 0 \}.
\]

When \( l(t) = t \), shortfall risk becomes the expectation of \( X_k \). We define the empirical estimate \( \hat{\rho}_{k,n}^S \) of the shortfall of arm \( k \) as

\[
\hat{\rho}_{k,n}^S \triangleq \inf \left\{ \kappa \in \mathbb{R} : \frac{1}{|S_k(n)|} \sum_{t \in S_k(n)} l(x_{k,t} - \kappa) \leq 0 \right\}.
\]
and the empirical estimate $\hat{\rho}_{\varphi,n}^S$ of the shortfall for the policy may be defined as

$$\hat{\rho}_{\varphi,n}^S \triangleq \inf \left\{ \kappa \in \mathbb{R} : \frac{1}{n} \sum_{t=1}^{n} l(x_{t\varphi,t} - \kappa) \leq 0 \right\}.$$  

The pseudo regret under the shortfall up to time $n$ is then

$$R_n^S(\varphi) \triangleq \mathbb{E} \inf \left\{ \kappa \in \mathbb{R} : \frac{1}{n} \sum_{t=1}^{n} l(x_{t\varphi,t} - \kappa) \leq 0 \right\} - SF_{k^*}.$$  

We make the following assumption about the underlying loss function $l$.

**Assumption 2.** The loss function $l(t)$ is continuous, strictly increasing and Lipschitz of order 1 in $t$ in the closed interval of $[-M, M]$ with Lipschitz constant $C_l$. In particular, we can let $l(0) = 0$ and $l(t)$ be bounded by a constant $M_l$ uniformly in $[0, M]$. Meanwhile, the derivative of $l(t)$ exists in $[-M, M]$ (with one-sided derivative at the boundary) and is uniformly lower bounded by a constant $m_l$, i.e. $l'(t) \geq m_l$ for all $t \in [-M, M]$.

Actually, Assumption 2 ensures that the set $\{ \kappa \in \mathbb{R} : \mathbb{E}[l(X_k - \kappa)] \leq 0 \}$ is non-empty for any arm and that $\rho_k^S$ lies within $[0, M]$, $k = 1, \ldots, K$ (as we will show in Lemma 7 below). Assumption 2 holds for many common loss functions including $l(t) = t$ which is the special case of expectation and $l(t) = e^t - 1$ which is the widely used exponential loss function. The term $\varepsilon_{\rho}(n, T_k(n), K, \delta)$ for shortfall is defined as follows

$$\varepsilon_{\rho}^S(n, T_k(n), K, \delta) \triangleq 2M_l M_G \sqrt{\frac{\log 4n^2K}{2T_k(n)}}.$$  

where $M_G$ is a constant which will be specified in Lemma 10. $\varepsilon_{\rho}^S$ is a direct confidence bound on the empirical shortfall risk of arm $k$. The index for arm $k$ for the shortfall case at time $n$ is then

$$B_{k,n}^S \triangleq \rho_{k,n}^S - \varepsilon_{\rho}^S(n, T_k(n), K, \delta).$$  

The upper bound for the pseudo regret under the shortfall risk is given in the following theorem.

**Theorem 3.** For all $n > K$, the pseudo regret $\tilde{R}_n^{SF}(\varphi)$ satisfies

$$R_n^S(\varphi) \leq (1 - \frac{\delta}{n})(\sum_{k \neq k^*} 8M_l^2 M_G \frac{\log 4n^2K}{\delta}) + 2M_l M_G \sqrt{\frac{\log 4n^2K}{2T_k^+(n)}} + \frac{\delta}{n} M,$$

where $\Delta_k^S \triangleq \rho_k^S - \rho_{k^*}^S$. Specifically, the order of the upper bound is dominated by the term $2M_l M_G \sqrt{\log (4n^2K/\delta)}[2T_k^+(n)]^{-1}$ which is of the order $O(\sqrt{\log n/n})$.

### 3.4 Discussion

For all three risk measures, we obtain a uniform upper bound of the order $O(\sqrt{\log n/n})$ on the pseudo regret. This convergence rate is different from the order of $O(\log n/n)$ in the risk-neutral case. In the risk-neutral MAB, the pseudo regret $\sum_{k=1}^{K} \mathbb{E}T_k(n)(\mu_{k^*} - \mu_k)$ is actually a linear
function with respect to the average number of pulls from the sub-optimal arms, and is thus dominated by the order of $E T_k(n)$. However, this situation does not hold for the risk-averse case. We explain in more detail for the specific cases of CVaR and MD. We reorganize the pseudo regret for CVaR and MD in a new consistent form, of which the derivation follows from the results in later sections. In this new form, the linear part and the non-linear part w.r.t $E T_k(n)$ are separated for clarity. The form is presented as below

$$\tilde{R}_n(\varphi) = \frac{1}{n} \sum_{k \neq k^*} E T_k(n) S_k + S^*.$$  

(4)

In the case of the CVaR risk measure,

$$S_k = (1 - \alpha)^{-1} E \left[ (X_k - \eta_{\varphi,n}) - (X_{k^*} - \eta_{\varphi,n})_+ \right],$$
$$S^* = (1 - \alpha)^{-1} E \left[ (X_{k^*} - \eta_{\varphi,n})_+ - (X_{k^*} - \eta^*)_+ \right] + E (\eta_{\varphi,n} - \eta^*),$$

when $\alpha = 0$, all $\eta$’s become the essential infimum. Therefore,

$$S_k = E [(X_k - \eta_{\varphi,n}) - (X_{k^*} - \eta_{\varphi,n})] = E (X_k - X_{k^*}),$$
$$S^* = E [(X_{k^*} - \eta_{\varphi,n})_+ - (X_{k^*} - \eta^*_+)] + E (\eta_{\varphi,n} - \eta^*) = 0.$$

In the case of the MD risk measure,

$$S_k = \Delta_k + \gamma |E |x_{k,t} - \mu_{\varphi,n}|^p - E |X_{k^*} - \mu_{k^*}|^p|,$$
$$S^* = \frac{\gamma E T_{k^*}(n)}{n} |E |X_{k^*} - \mu_{\varphi,n}|^p - E |X_{k^*} - \mu_{k^*}|^p|,$$

when $\gamma = 0$,

$$S_k = \Delta_k = E (X_k - X_{k^*}),$$
$$S^* = 0.$$

In the risk-averse case, the pseudo regret includes the nonlinear term of $S^*$. As each $S_k$ is bounded, the first term of (4) is dominated by the order of $\frac{1}{n} \sum_{k \neq k^*} E T_i(n)$, which is of the order $O(\log n/n)$. However, the second term is dominated by the asymptotic order of the specific statistic in each case. As a result, the overall order of the upper bound is actually $O(\sqrt{\log n/n})$.

In the special case where the three risk measures reduce to expectation, the second term $S^*$ is zero and the risk-averse pseudo regret is consistent with the risk-neutral pseudo regret. Thus, the asymptotic order of the upper bound on the pseudo regret becomes $O(\log n/n)$ for the risk-neutral case as expected.

4 Proofs of the main results

In this section, we give the detailed proofs of our three main theorems. Throughout this section, we use the sequence $\{x_{k,t}\}_{t=1}^n$ which is an i.i.d. sequence with each $x_{k,t}$ subject to $F_k$. We introduce this sequence for notational convenience. Note that $\{x_{k,t}\}_{t=1}^n$ is not actually chosen
by any algorithm, thus is unobservable.

The proofs in Sections 4.1-4.3 all follow a similar general procedure. Recall that the index for each arm consists of two parts: $\hat{\rho}_{k,n}$ and $\varepsilon_{\rho}(\cdot)$. We first prove that $\varepsilon_{\rho}(\cdot)$ is an upper bound on the estimation error $|\hat{\rho}_{k,n} - \rho_k|$ with a probability of $1 - O(n^{-2})$. Thus, the total probability that all $\rho_k$’s falls below the index at each of the first $n$ times is of order $O(n^{-1})$, which means a sub-optimal arm becomes less likely to be chosen by the algorithm as the number of pulls of it increases. Based on this conclusion, we may further show that the total number of pulls of all sub-optimal arms is upper bounded. This means that the portion of pulls of the optimal arm in the whole sequence increases, thus, the empirical risk based on this sequence is supposed to approximate the true risk of the optimal arm. Finally, we may prove the upper bound for the pseudo regret in each of our three main theorems. The proof of each starred lemma is detailed in the appendix.

### 4.1 Conditional value-at-risk measure

Recall the definitions of $\eta_k$, $\eta_{k,n}$, $\eta_{\varphi,n}$, $\eta^*$ and $\eta^*_n$ in Section 3.1. In this section, we first derive a probability upper bound for $|\eta_{k,n} - \eta_k|$ in Lemma 2. Based on the results in Lemma 2, we may establish an upper bound on $\hat{\rho}_{k,n}^C$, then we may bound the number of pulls of all sub-optimal arms in Lemma 4. In Lemma 5, we construct the upper bound for another crucial term $|\eta_{\varphi,n} - \eta^*|$ that appears in the pseudo regret. Based on the results in Lemmas 2-5, we may decompose the pseudo regret into different parts and bound each part. Then we prove the upper bound for the pseudo regret in Theorem 1 in the end of this section.

**Step 1: Bounding $|\eta_{k,n} - \eta_k|$ which leads to a bound on $|\hat{\rho}_{k,n}^C - \rho_k^C|$.

**Lemma 2.** Suppose Assumption 1 holds, then for all $k = 1, \ldots, K$, we have

$$\eta_{k,n} \xrightarrow{P} \eta_k,$$

as $n \to \infty$, and $\forall n > 1$, with probability at least $1 - \delta$

$$|\eta_{k,n} - \eta_k| \leq \frac{2}{f_k(\eta_k)} \sqrt{\log \frac{4}{\delta} \frac{1}{2n}}. \tag{6}$$

Furthermore, noting $M$ is the uniform essential bound on all $X_k$’s, we have

$$\mathbb{E}(|\eta_{k,n} - \eta_k|) \leq \frac{2}{f_k(\eta_k)} \sqrt{\log \frac{4}{\delta} \frac{1}{2n}} + \delta M. \tag{7}$$

**Proof.** Note that $\eta_{k,n}$ is the $\alpha$-quantile of the empirical distribution of arm $k$, i.e. the sample percentile of $\{x_{k,t}\}_{t=1}^n$. From the Bahadur representation of sample percentiles ([DasGupta 2008] Theorem 17.1), we have

$$\eta_{k,n} - \eta_k = \frac{1}{n} \sum_{t=1}^n \frac{\alpha - I\{x_{k,t} \leq \eta_k\}}{f_k(\eta_k)} + r_n, \tag{8}$$

where $r_n = o_p(n^{-1/2})$. Since $\mathbb{E}(\alpha - I\{x_{k,t} \leq \eta_k\}) = 0$, by the strong law of large numbers,
\[ \eta_{k,n} - \eta_k \xrightarrow{P} 0. \] Meanwhile, \( \alpha - I \{ x_{k,t} \leq \eta_k \} \in [\alpha - 1, \alpha] \), so by Hoeffding’s inequality, we have

\[
P \left\{ \left| \frac{1}{n} \sum_{t=1}^{n} \frac{\alpha - I \{ x_{k,t} \leq \eta_k \}}{f_k(\eta_k)} \right| \geq \frac{1}{f_k(\eta_k)} \sqrt{\log \frac{4}{\delta}} \right\} \leq \frac{\delta}{2}. \tag{9} \]

Combining (8) with (9), we see that

\[
P \left\{ |\eta_{k,n} - \eta_k| \geq \frac{2}{f_k(\eta_k)} \sqrt{\log \frac{4}{\delta}} \right\} \leq \frac{2}{f_k(\eta_k)} \sqrt{\log \frac{4}{\delta}} + \frac{\delta}{2} \leq \delta.
\]

Conclusion (6) follows by noting that the term \( o(1) \) in the inequality above is actually smaller than 1. Conclusion (7) follows by noting that \( |\eta_{k,n} - \eta_k| \) is bounded by \( M \). \hfill \Box

**Step 2: Bounding the number of pulls of each sub-optimal arm.** To further analyze the upper bound, we define a key subset as follows

\[
\Omega_n^C \triangleq \left\{ \forall k = 1, \ldots, K, \forall t = 1, \ldots, n, \quad |\eta_{k,t} - \eta_k| \leq 2m(\alpha) \sqrt{\log \frac{2n \delta K}{2t}} \right\},
\]

where \( m(\alpha) \) has been defined in Assumption 1. The subset \( \Omega_n^C \) consists of all the realizations that each \( \eta_{k,t} \) is close enough to its real value \( \eta_k \) at each of the first \( n \) times. Confined in \( \Omega_n^C \), the probability that the index of any sub-optimal arm is larger than that of the optimal arm is controlled, with an lower bound given in the following lemma.

**Lemma 3.** For all \( n \geq 1 \), we have \( P\{\Omega_n^C\} \geq 1 - 3\delta/n \).

Through the proofs of Lemma 2 and 3, we have shown that \( \varepsilon^C_\rho \) satisfies Condition 1. In the next lemma, we construct an upper bound on the number of pulls of each sub-optimal arm.

**Lemma 4.** For any sub-optimal arm \( k \neq k^* \) and for all \( n \geq 1 \), the following relation holds with probability at least \( 1 - 3\delta/n \)

\[
T_k(n) \leq 2 \log \frac{2n^2K}{\delta} \left( \frac{2 \left[ 1 + (1 - \alpha)^{-1} \right] m(\alpha) + (1 - \alpha)^{-1} M}{\Delta_k^{\text{CVaR}} - (1 - \alpha)^{-1} \frac{12}{n^2K} M} \right)^2. \tag{10} \]
Furthermore,

$$\mathbf{E}T_k(n) \leq (1 - 3\delta)2 \log \frac{2n^2K}{\delta} \left(2 \left[1 + (1 - \alpha)^{-1}\right] m(\alpha) + (1 - \alpha)^{-1} M \right)^2 + 3\delta,$$  \hspace{1cm} (11)$$

specifically, $$\mathbf{E}T_k(n) \leq O(\log n).$$

**Proof.** The main idea of the proof is to upper bound the difference between the empirical risk and the real risk of each arm. Further, constrained to $$\Omega_n^C$$, both an upper bound and a lower bound are available for the index of each arm. From this point, we may bound the number of pulls of each sub-optimal arm using the relation between the indices of the corresponding sub-optimal arm and the optimal arm. We first decompose $$\hat{\rho}_{k,n}^C$$ as

$$\hat{\rho}_{k,n}^C = \eta_{k,T_k(n)} + (1 - \alpha)^{-1} \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} (x_{k,t} - \eta_{k,T_k(n)}) +$$

$$+ (1 - \alpha)^{-1} \left\{ \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} (x_{k,t} - \eta_{k,T_k(n)}) + - \mathbf{E}[(X_k - \eta_{k,T_k(n)})^+] \right\}$$

$$+ (1 - \alpha)^{-1} \{ \mathbf{E}[(X_k - \eta_{k,T_k(n)})^+] - \mathbf{E}[(X_k - \eta_k)^+] \}.$$

We will bound the last three terms in the right side of the equation above. The last term in the above equation is bounded below by

$$- (1 - \alpha)^{-1} \mathbf{E} (|\eta_{k,T_k(n)} - \eta_k|)$$

$$\leq (1 - \alpha)^{-1} \{ \mathbf{E} [(X_k - \eta_{k,T_k(n)})^+] - \mathbf{E} [(X_k - \eta_k)^+] \}$$

$$\leq (1 - \alpha)^{-1} \mathbf{E} (|\eta_{k,T_k(n)} - \eta_k|).$$  \hspace{1cm} (12)$$

Let $$2\delta/n^2K$$ replace $$\delta$$ in Lemma 2, we further obtain the following lower and upper bounds

$$- (1 - \alpha)^{-1} \left( 2M(\alpha) \sqrt{\frac{\log \frac{2n^2K}{\delta}}{2T_k(n)}} + \frac{2\delta}{n^2K} M \right)$$

$$\leq (1 - \alpha)^{-1} \{ \mathbf{E} [(X_k - \eta_{k,T_k(n)})^+] - \mathbf{E} [(X_k - \eta_k)^+] \}$$

$$\leq (1 - \alpha)^{-1} \left( 2m(\alpha) \sqrt{\frac{\log \frac{2n^2K}{\delta}}{2T_k(n)}} + \frac{2\delta}{n^2K} M \right).$$  \hspace{1cm} (13)$$

Meanwhile, restricted to the subset $$\Omega_n^C$$ and for $$t = T_k(n)$$, it is true that

$$- 2m(\alpha) \sqrt{\frac{\log \frac{2n^2K}{\delta}}{2T_k(n)}} \leq \eta_{k,T_k(n)} - \eta_k \leq 2m(\alpha) \sqrt{\frac{\log \frac{2n^2K}{\delta}}{2T_k(n)}},$$  \hspace{1cm} (14)$$
and

\[-(1 - \alpha)^{-1} M \sqrt{\frac{\log 2n^2 K}{2T_k(n)}} \leq (1 - \alpha)^{-1} \left\{ \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} (x_{k,t} - \eta_{k,T_k(n)}) + -E[(X_k - \eta_k, T_k(n))^-] \right\} \]

\[\leq (1 - \alpha)^{-1} M \sqrt{\frac{\log 2n^2 K}{2T_k(n)}}. \tag{15}\]

Combining (12)-(15), we see that \(\hat{\rho}^C_{k,n}\) is bounded. Then we conclude that the following relation holds for the index for any arm \(k, k = 1, \ldots, K\) with probability at least \((1 - 3\delta/n)\), in particular, for the optimal arm \(k^*\)

\[\rho_k^C - 2 \left( 2 \left[ 1 + (1 - \alpha)^{-1} \right] m(\alpha) + (1 - \alpha)^{-1} M \right) \sqrt{\frac{\log 2n^2 K}{2T_k(n)}} - (1 - \alpha)^{-1} \frac{2\delta}{n^2 K} M \]

\[\leq B_{k,n}^C \leq \rho_k^C + (1 - \alpha)^{-1} \frac{2\delta}{n^2 K} M. \]

At any time \(n + 1\), if arm \(k \neq k^*\) is pulled, i.e. \(I_n = k\), then the index of arm \(k\) must be lower than that of any other arm, including the optimal arm \(k^*\). So we must have

\[B_{k,n}^C = \rho_k^C - (1 - \alpha)^{-1} \left( 1 - \frac{3\delta}{n} \right) M \sqrt{\frac{\log 2n^2 K}{2T_k(n)}} \]

\[\leq \rho_{k^*,n}^C - (1 - \alpha)^{-1} \left( 1 - \frac{3\delta}{n} \right) M \sqrt{\frac{\log 2n^2 K}{2T_k(n)}} = B_{k^*,n}^C, \]

which means, with probability at least \((1 - 3\delta/n)\), the following relation holds

\[\rho_k^C - 2 \left( 2 \left[ 1 + (1 - \alpha)^{-1} \right] m(\alpha) + (1 - \alpha)^{-1} M \right) \sqrt{\frac{\log 2n^2 K}{2T_k(n)}} \]

\[- (1 - \alpha)^{-1} \frac{2\delta}{n^2 K} M \leq \rho_{k^*,n}^C + (1 - \alpha)^{-1} \frac{2\delta}{n^2 K} M. \tag{16}\]

Relation (16) directly leads to conclusion (10). Conclusion (11) follows by noting that \(T_k(n)\) is automatically bounded by \(n\). \(\Box\)

**Step 3: Establishing an upper bound on the pseudo regret.** The convergence of the sequence \(\{\eta_{\varphi,n}\}_{n=1}^{\infty}\) is necessary for our derivation of the upper bound on the pseudo regret. Under general conditions, there is no guarantee of the convergence of the sequence \(\eta_{\varphi,n}\) as \(n \to \infty\). However, if we let \(k_n \triangleq \sum_{k \neq k^*} T_k(n)\) be the total number of pulls from sub-optimal arms, then we see that \(k_n = O(\log n)\). Based on this fact, the following lemma establishes the convergence of the sequence \(\{\eta_{\varphi,n}\}_{n=1}^{\infty}\).

**Lemma 5.** If \(k_n\) satisfies \(k_n = O(\log n)\), then

\[\eta_{\varphi,n} \xrightarrow{p} \eta^*, \tag{17}\]
as \( n \to \infty \), and \( \forall n \geq 1 \), with probability at least \( 1 - \delta \) we have

\[
|\eta_{\varphi,n} - \eta^*| \leq \max \{ \alpha, 1 - \alpha \} O \left( \frac{\log n}{n} \right) m(\alpha) + 2m(\alpha) \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\]  

(18)

**Proof.** First we separate the observation sequence up to time \( n \) into two parts: \( k_n \) observations from the sub-optimal arms and \( (n - k_n) \) observations from the optimal arm. By considering the extreme cases where all the observations in the latter category are larger or smaller than those in the former, we conclude that

\[
F_{k^*,(n-k_n)}^{-1} \left[ \frac{1}{\alpha} - \frac{k_n}{n} \right] \leq \eta_{\varphi,n} \leq F_{k^*,(n-k_n)}^{-1} \left[ \frac{1}{\alpha} \right].
\]

Let \( \alpha_1 = (1 + k_n/n) \alpha - k_n/n \) and \( \alpha_2 = (1 + k_n/n) \alpha \). Based on Lemma 2, for all \( n \geq 1 \), the following bounds on \( \eta_{\varphi,n} \) hold with probability at least \( 1 - \delta \)

\[
F_{k^*,(n-k_n)}^{-1}(\alpha_1) - \frac{2}{f_{k^*}(F_{k^*,(n-k_n)}^{-1}(\alpha_1))} \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \leq \eta_{\varphi,n} \leq F_{k^*,(n-k_n)}^{-1}(\alpha_2) + \frac{2}{f_{k^*}(F_{k^*,(n-k_n)}^{-1}(\alpha_2))} \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\]

Considering \( \eta^* = F_{k^*,1}^{-1}(\alpha) \), the following bound on \( |\eta_{\varphi,n} - \eta^*| \) holds with probability \( 1 - \delta \)

\[
|\eta_{\varphi,n} - \eta^*| \leq \max \left\{ F_{k^*,1}^{-1}(\alpha_1) - F_{k^*,1}^{-1}(\alpha_2) + \frac{2}{f_{k^*}(F_{k^*,1}^{-1}(\alpha_2))} \sqrt{\frac{\log \frac{1}{\delta}}{2n}}, F_{k^*,1}^{-1}(\alpha_2) - F_{k^*,1}^{-1}(\alpha_1) + \frac{2}{f_{k^*}(F_{k^*,1}^{-1}(\alpha_1))} \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right\}.
\]

Under Assumption 1, from the mean value theorem and the condition \( k_n = O(\log n) \), as \( n \to \infty \) we have

\[
F_{k^*,1}^{-1}(\alpha_1) - F_{k^*,1}^{-1}(\alpha_2) = O\left( \frac{k_n}{n} \right)(1 - \alpha) \frac{1}{f_k(\eta_k)} \leq O\left( \frac{k_n}{n} \right)(1 - \alpha)m(\alpha),
\]

and

\[
F_{k^*,1}^{-1}(\alpha_2) - F_{k^*,1}^{-1}(\alpha_1) \leq O\left( \frac{\log n}{n} \right) \alpha m(\alpha).
\]

The desired conclusion then follows. \( \square \)

With upper confidence bounds on both \( |\eta_{\varphi,n} - \eta^*| \) and \( T_k(n) \), we can now derive the upper bound on the pseudo regret \( R^C_{\eta_{\varphi}}(\varphi) \) in Theorem 1.

**Proof of Theorem 1.** Based on Lemma 3, restricted to the set \( \Omega_{n}^C \), \( |\eta_{\varphi,n} - \eta^*| \) can be upper bounded by \( M_{\varphi}(n) \) with probability at least \( 1 - \delta/n \). Define the subset by \( \Lambda_n \subseteq \Omega_{n}^C \) where \( |\eta_{\varphi,n} - \eta^*| \) is bounded, then \( P\{\Lambda_n \} \geq 1 - 4\delta/n \). So with probability at least \( (1 - \delta/n)P\{\Omega_{n}^C \} = (1 - \delta/n)(1 - 3\delta/n) \geq 1 - 4\delta/n \), \( |\eta_{\varphi,n} - \eta^*| \) can be upper bounded by \( M_{\varphi}(n) \). We first decompose
and bound the pseudo regret with

\[
R^C_n(\varphi) \leq E(|\eta_{\varphi, n} - \eta^*||\Lambda_n) + E(|\eta_{\varphi, n} - \eta^*||\Lambda^c_n)
\]

\[
+ (1 - \alpha)^{-1} \frac{1}{n} \sum_{k=1}^{K} ET_k(n)E[(X_k - \eta_{\varphi, n})_+ - (X_k^* - \eta^*)_+]|A_n|
\]

\[
+ (1 - \alpha)^{-1} \frac{1}{n} \sum_{k=1}^{K} ET_k(n)E[(X_k - \eta_{\varphi, n})_+ - (X_k^* - \eta^*)_+]|\Lambda^c_n].
\] (19)

Note that

\[
E [(X_k - \eta_{\varphi, n})_+ - (X_k^* - \eta^*)_+] \leq E [(X_k - \eta_k)_+] - E [(X_k^* - \eta^*)_+] + |\eta_{\varphi n} - \eta_k|
\]

\[
\leq \Delta^C_k + M.
\]

Meanwhile, since \(|\eta_{\varphi, n} - \eta^*|\) is naturally bounded by \(M\) and \(ET_k(n)\) is bounded by \(M_k(n)\), each term in (19) can now be bounded. Then, we have

\[
R^C_n(\varphi) \leq (1 - \frac{4\delta}{n})M_{\varphi}(n) + \frac{4\delta}{n} M + (1 - \alpha)^{-1} (1 - \frac{4\delta}{n}) \frac{1}{n} \sum_{k \neq k^*} M_k(n)(\Delta^C_k + M)
\]

\[
+ (1 - \alpha)^{-1} \frac{4\delta}{n} (\Delta^C + M).
\]

The conclusion then follows from combining the terms in the above inequality together. □

4.2 Mean-deviation risk measure

In this section, we first give some preliminary results in Lemma 6 and Lemma 7. Then in Lemma 9, we derive the upper bound on the number of pulls of each sub-optimal arm. In the end of the section, we give the proof of Theorem 2.

Step 1: Preliminary results. We begin with the following lemma.

**Lemma 6.** Let \(f(x) = x^p\) and \(g(x) = x^{\frac{1}{p}}\) be defined on \([0, +\infty)\), where \(p \geq 1\). Then, for any \(x, y \in [0, +\infty)\), we have

\[
|f(x) - f(y)| \leq f'(\max\{x, y\}) |x - y|,
\]

\[
|g(x) - g(y)| \leq g(|x - y|),
\]

specifically, if \(x, y \in [0, M]\), then \(|f(x) - f(y)| \leq pM^{p-1} |x - y|\).

The conclusion in Lemma 6 follows directly from the convexity of \(f(x)\) and the concavity of \(g(x)\). The proof is trivial so we omit it. We also need the preliminary result in the following lemma.

**Lemma 7.** For \(x \geq 0\), if

\[
a x + b x^{\frac{1}{p}} \geq c,
\]
where \(a, b\) and \(c\) are fixed constants, \(p \geq 1\), then we can conclude that
\[
x \geq \min \left\{ 1, \frac{c}{a+b} \left( \frac{c}{a+b} \right)^p \right\}.
\]

**Step 2: Bounding the number of pulls of each sub-optimal arm.** For further discussion, we define the following key subset
\[
\Omega^M_n \triangleq \left\{ \forall i = k, \ldots, K, \forall t = 1, \ldots, n, \text{s.t.} |\overline{x}_{k,t} - \mu_k| \leq M \sqrt{\frac{\log 4n^2K}{2t}} \right\}.
\]

The subset \(\Omega^M_n\) has the same meaning as \(\Omega^C_n\), which represents the set of realizations where the empirical risk of each arm is close enough to the real risk at each of the first \(n\) times. By establishing the lower bound of the probability of \(\Omega^M_n\) in the following lemma, we may show that the \(\hat{\rho}^M_{k,n}\) approximates \(\rho^M_k\) with a high probability.

**Lemma 8.** For all \(n \geq 1\), we have \(\Pr\{\Omega^M_n\} \geq 1 - \delta/n\).

From the proof of Lemma 8, we have shown that \(\hat{\rho}^M\) satisfies Condition 1. In the following lemma, we bound the number of pulls from the sub-optimal arms. We let
\[
m_k \triangleq \min\left\{1, \frac{\rho^M_k - \hat{\rho}^M_k}{2M[1 + (1 + p)^{1/p}]}, \frac{(\rho^M_k - \hat{\rho}^M_k)^p}{2pM^p[1 + (1 + p)^{1/p}]^p} \right\}^{-2},
\]
for all \(k = 1, \ldots, K\).

**Lemma 9.** For any sub-optimal arm \(k \neq k^*\), the following relation holds with probability at least \(1 - \delta/n\)
\[
T_k(n) \leq m_k \log \frac{4n^2K}{\delta}.
\]
Furthermore,
\[
\mathbb{E}T_k(n) \leq \left(1 - \frac{\delta}{n}\right) m_k \log \frac{4n^2K}{\delta} + \delta.
\]

Specifically, we have \(\mathbb{E}T_i(n) \leq O(\log n)\).

**Proof.** The main idea of the proof is the same as Lemma 4. Our task is to establish the upper and lower bounds for the index of arm \(k\) based on the bound of \(|\hat{\rho}^M_{k,n} - \rho^M_k|\), for all \(k = 1, \ldots, K\).

The empirical estimate \(\hat{\rho}^M_{k,n}\) consists of two terms, we first bound these two terms in this proof. Restricted to the set \(\Omega^M_n\), we have
\[
-M \sqrt{\frac{\log 4n^2K}{2T_k(n)}} \leq \overline{x}_{k,T_k(n)} - \mu_k \leq M \sqrt{\frac{\log 4n^2K}{2T_k(n)}}.
\]
So we see that the first term of \(\hat{\rho}^M_{k,n}\) is bounded. From Lemma 6, we have
\[
\left( \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} |x_{k,t} - \overline{x}_{k,T_k(n)}|^p \right)^{\frac{1}{p}} - \|X_k\|_p \leq \left( \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} |x_{k,t} - \overline{x}_{k,T_k(n)}|^p \right)^{\frac{1}{p}} - \mathbb{E}|X_k - \mu_k|^p.
\]
The right side of the inequality above is further bounded by

\[
\frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} |x_{k,t} - \overline{x}_{k,T_k(n)}|^p - E|X_k - \mu_k|^p \leq \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} |x_{k,t} - \mu_k|^p - |x_{k,t} - \mu_k|^p \leq pM^{p-1} \left( |x_{k,t} - \overline{x}_{k,T_k(n)}| - |x_{k,t} - \mu_k| \right) + M^{p} \sqrt{\log \frac{4n^2K}{\delta}} 2T_k(n) \leq \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} \left( |x_{k,t} - \overline{x}_{k,T_k(n)}|^p - |x_{k,t} - \mu_k|^p \right) + \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} \left( |x_{k,t} - \mu_k| \right) + M^{p} \sqrt{\log \frac{4n^2K}{\delta}} 2T_k(n) \leq (p + 1)M^{p} \sqrt{\log \frac{4n^2K}{\delta}} 2T_k(n),
\]

where the third inequality follows from the conclusion in Lemma 5 and the last inequality follows from the fact that \(|\overline{x}_{k,T_k(n)} - \mu_k| \leq M\). So, we have

\[-M[(p + 1) \sqrt{\frac{\log 4n^2K}{\delta}} 2T_k(n)]^{\frac{1}{p}} \leq \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} |x_{k,t} - \overline{x}_{k,T_k(n)}|^p \leq \frac{1}{T_k(n)} \sum_{t=1}^{T_k(n)} \left| |x_{k,t} - \mu_k| \right| \leq M[(p + 1) \sqrt{\frac{\log 4n^2K}{\delta}} 2T_k(n)]^{\frac{1}{p}}.\]

From the three relations above, we see that the second term of \(\hat{\rho}_{k,n}^M\) is also bounded. Suppose that arm \(k \neq k^*\) is pulled at time \(n + 1\) by Algorithm 1, then the following relation must hold

\[\rho_k^M - 2M \sqrt{\frac{\log 4n^2K}{T_k(n)}} - 2M \left( p + 1 \sqrt{\frac{\log 4n^2K}{2T_k(n)}} \right)^{\frac{1}{p}} \leq B_k^M, n \leq B_k^M, n \leq \rho_k^M,\]

and so

\[\rho_k^M - \rho_k^M \leq 2M \sqrt{\frac{\log 4n^2K}{T_k(n)}} + 2M \left( p + 1 \sqrt{\frac{\log 4n^2K}{2T_k(n)}} \right)^{\frac{1}{p}}.\]

As in the proof of Lemma 4, conclusion (20) follows by substituting \(x = \sqrt{T_k^{-1}(n)\log(4n^2K/\delta)}, a = 2M, b = 2M (p + 1)^{\frac{1}{p}}\) and \(c = \rho_k^M - \rho_k^M\) in Lemma 7. Conclusion (21) directly follows by noting \(T_k(n)\) is automatically bounded by \(n\). \qed

**Step 3: Establishing an upper bound on the pseudo regret.** Note that the pseudo regret consists of two parts: the differences between means and deviations. Based on the previous two steps, we can bound each of these two parts. The derivation of the upper bound for the pseudo regret \(R_n^M(\varphi)\) in Theorem 2 is given below.
Proof of Theorem 2. We first decompose the pseudo regret as follows

\[ R_n^M(\varphi) = \left( E \left( \mathbf{f}_n, \varphi - X^* \right) \right) + \gamma \left( \frac{1}{n} \sum_{t=1}^{n} \left( x_{t, t} - \mathbf{f}_n, \varphi - X^* \right) \right) \]

\[ \leq \frac{1}{n} \sum_{k \neq k^*} E T_k(n) \Delta^M_{k^*} + \gamma \left( \frac{1}{n} \sum_{t=1}^{n} E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right) \]

\[ \leq \frac{1}{n} \sum_{k \neq k^*} E T_k(n) \Delta^M_{k^*} + \gamma \left( \frac{1}{n} \sum_{t=1}^{n} E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right) \]

\[ \leq \frac{1}{n} \sum_{k \neq k^*} E T_k(n) \Delta^M_{k^*} + \frac{\gamma}{n} \sum_{k=1}^{K} E T_k(n) \left| E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right| \]

\[ = \frac{1}{n} \sum_{k \neq k^*} E T_k(n) \Delta^M_{k^*} + \frac{\gamma}{n} \sum_{k \neq k^*} E T_k(n) \left| E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right| \]

\[ + \frac{\gamma}{n} \sum_{k \neq k^*} E T_k(n) \left| E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right| , \tag{22} \]

where the second inequality comes from Lemma 6. Next, we bound the last two terms in (22). The second term in (22) is bounded by

\[ \frac{\gamma}{n} \sum_{k \neq k^*} E T_k(n) \left| E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right| \leq \frac{\gamma p M^p}{n} \sum_{k \neq k^*} E T_k(n), \]

due to Lemma 6. The third term in (22) is bounded by

\[ \frac{\gamma}{n} \sum_{k \neq k^*} E T_k(n) \left| E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right| \leq \frac{\gamma p M^p}{n} \left| E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right| \]

\[ \leq \frac{\gamma p M^p}{n} \left| E \left| x_{t, t} - \mathbf{f}_n, \varphi - X^* \right| \right| \]

Furthermore, the first term above is bounded by \[ \frac{\gamma p M^p}{n} \sum_{k \neq k^*} E T_k(n) \]. Restricted to the subset \[ \Omega_n^M \], the second term above satisfies the bound

\[ \left| \frac{1}{T_{k^*}(n)} \sum_{t=1}^{T_{k^*}(n)} \left( x_{k^*, t} - \mathbf{f}_n, \varphi - X^* \right) \right| = \left| \mathbf{f}_n, T_{k^*}(n) - X^* \right| \leq M \sqrt{\log \frac{4n^2 K^2}{\alpha}} \]
Meanwhile, \(|x_{k^*}, T_{k^*}(n) - \mu_{k^*}|\) is bounded by \(M\), and so
\[
E \left| \frac{1}{T_{k^*}(n)} \sum_{t=1}^{T_{k^*}(n)} (x_{k^*,t} - \mu_{k^*}) \right| \leq \left( 1 - \frac{\delta}{n} \right) M \sqrt{\frac{\log \frac{4n^2K}{\delta}}{2T_{k^*}(n)}} + \frac{\delta}{n} M.
\]
From this reasoning, we see that both the last two terms in (22) are bounded which gives the desired conclusion.

4.3 Shortfall risk measure

In Lemmas 10-12, we first establish that \(\hat{\rho}^S_{k,n}\) is an M estimator [(DasGupta 2008) Definition 17.1] of \(\rho_k^S\). Then we derive an upper bound for \(|\hat{\rho}^S_{k,n} - \rho_k^S|\), based on which, we establish the upper bound on the number of pulls of each sub-optimal arm. Then we prove Theorem 3 in the end of the section.

**Step 1: Bounding \(|\hat{\rho}^S_{k,n} - \rho_k^S|\) by the concentration results for M estimators.** Let \(G_k(\kappa) \triangleq E[l(X_k - \kappa)]\) for notational convenience, based on Assumption 2, we conclude that \(G_k(\kappa)\) obeys the following property.

**Lemma 10.** The derivative of \(G_k(\kappa)\), denoted by \(G_k'(\kappa)\), exists in \((0, M)\) and \([G_k'(SF_k)]^{-1}\) is upper bounded by a constant \(M_G\), for all \(k = 1, \ldots, K\).

The following lemma establishes that \(\hat{\rho}^S_{k,n}\) is an M estimator and converges to \(\rho_k^S\) a.s.

**Lemma 11.** Suppose Assumption 2 holds, \(\hat{\rho}^S_{k,n}\) is the unique solution in \(\kappa\) of
\[
\int_0^M l(X - \kappa)dF_{i,n}(X) = 0.
\]
Then \(\hat{\rho}^S_{k,n}\) is an M estimator of \(\rho_k^S\). Further, \(\rho_k^S\) is the unique solution of \(G_k(\rho_k^S) = 0\), and \(\hat{\rho}^S_{k,n} \overset{a.s.}{\to} \rho_k^S\).

Based on the conclusion in Lemma 11, we may obtain the Bahadur representation for \(|\hat{\rho}^S_{k,n} - \rho_k^S|\). Subsequently, we may establish the confidence bound on \(\hat{\rho}^S_{k,n}\) which is given in the following lemma.

**Lemma 12.** For all \(n \geq 1\), the following bound on \(|\hat{\rho}^S_{k,n} - \rho_k^S|\) holds with probability at least \(1 - \delta\)
\[
|\hat{\rho}^S_{k,n} - \rho_k^S| \leq 2M_1M_G\sqrt{\frac{\log \frac{4}{\delta}}{2n}}. \quad (23)
\]
Furthermore,
\[
E \left( |\hat{\rho}^S_{k,n} - \rho_k^S| \right) \leq 2M_1M_G\sqrt{\frac{\log \frac{4}{\delta}}{2n}} + \delta M. \quad (24)
\]

**Proof.** From the conclusion of Lemma 11, \(\hat{\rho}^S_{k,n}\) is an M estimator that converges to \(\rho_k^S\). The Bahadur representation for \(\rho_k^S - \hat{\rho}^S_{k,n}\) is given as below [(DasGupta 2008) Theorem 17.3]
\[
\rho_k^S - \hat{\rho}^S_{k,n} = \frac{1}{n} \sum_{t=1}^{n} \frac{l(x_{k,t} - \rho_k^S)}{G_k'(\rho_k^S)} + r_n,
\]
where \( r_n = o_p(1/\sqrt{n}) \). Noting that \( \mathbb{E}[l(x_{k,t} - \rho^S_k)] = 0 \) and each \( l(x_{k,t} - \rho^S_k)[G'_k(\rho^S_k)]^{-1} \) is bounded by \( M_l M_G \), by Hoeffding’s inequality we have

\[
P \left\{ \left| \frac{1}{n} \sum_{t=1}^{n} l \left( x_{k,t} - \rho^S_k \right) \right| \geq M_l M_G \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \right\} \leq \frac{\delta}{2}.
\]

Then, we have the following bound on the probability that \( |\hat{\rho}_{k,n}^S - \rho^S_k| \) exceeds a certain threshold

\[
P \left\{ |\hat{\rho}_{k,n}^S - \rho^S_k| \geq 2M_l M_G \sqrt{\frac{\log \frac{4n^2K}{2t}}{2t}} \right\} \leq \delta.
\]

Conclusion (23) follows from this bound. As \( \hat{\rho}_{k,n}^S \in [0, M] \) and \( \rho^S_k \in [0, M] \), we know that \( |\hat{\rho}_{k,n}^S - \rho^S_k| \leq M \), and so conclusion (24) follows.

Lemma 12 establishes an probability upper bound on \( |\hat{\rho}_{k,n}^S - \rho^S_k| \), based on which, we can establish the index policy for the shortfall case.

**Step 2: Bounding the number of pulls of each sub-optimal arm.** The following lemma gives bounds on \( T_k(n) \) and \( \mathbb{E}T_i(n) \) in probability. We define the following key subset

\[
\Omega^S_n \triangleq \left\{ \forall k = 1, \ldots, K, \forall t = 1, \ldots, n, \text{s.t. } |\hat{\rho}_{k,n}^S - \rho^S_k| \leq 2M_l M_G \sqrt{\frac{\log \frac{4n^2K}{2t}}{2t}} \right\}.
\]

As before, subset \( \Omega^S_n \) collects all the realizations where the empirical risk of each arm is close enough to its real risk at each of the first \( n \) times. The following lemma gives a lower bound on the probability of \( \Omega^S_n \).

**Lemma 13.** For all \( n \geq 1 \), we have \( \mathbb{P}\{\Omega^S_n\} \geq 1 - \delta/n \).

Through the proofs of Lemma 12 and Lemma 13, we have shown that \( \varepsilon^S \) satisfies Condition 1. In the following lemma, we establish an upper bound on the number of pulls of sub-optimal arms.

**Lemma 14.** For any sub-optimal arm \( k \neq k^* \), the following relation holds with probability at least \( 1 - \delta/n \)

\[
T_k(n) \leq \frac{8M_l M_G}{\Delta^S_k} \log \frac{4n^2K}{\delta}.
\]

Furthermore,

\[
\mathbb{E}T_k(n) \leq \left( 1 - \frac{\delta}{n} \right) \frac{8M_l M_G}{\Delta^S_k} \log \frac{4n^2K}{\delta} + \delta.
\]

Specifically, \( \mathbb{E}T_k(n) \leq O(\log n) \).

**Proof.** The main idea of the proof is the same as Lemma 4. As \( |\hat{\rho}_{k,n}^S - \rho^S_k| \) has been bounded in Lemma 12, we can directly establish an upper bound on the number of pulls of each sub-optimal arm in probability using the relation of indices between each sub-optimal arm and the optimal arm. Restricted to the set \( \Omega^S_n \), the following inequality holds for each arm, including
the optimal arm $k^*$

$$-2M_lM_G \log \frac{4n^2K}{\delta} \leq \hat{\rho}^S_{k,n} - \rho^S_k \leq 2M_lM_G \log \frac{4n^2K}{2T_k(n)}.$$

At time $n + 1$, if any sub-optimal arm $k \neq k^*$ is pulled, then we must have

$$\rho^S_k - 4M_lM_G \log \frac{4n^2K}{\delta} \leq B^S_{k,n} \leq B^S_{k^*,n} \leq \rho^S_{k^*}.$$

Conclusion (25) then follows directly from the above inequality and conclusion (26) follows by noting that $T_k(n)$ is automatically bounded by $n$.

Step 3: Establishing an upper bound on the pseudo regret. Based on previous results we have upper bounds both on $|\hat{\rho}^S_{k,n} - \rho^S_k|$ and the number of pulls of sub-optimal arms. We may now give the full derivation of the upper bound on the pseudo regret in Theorem 3.

Proof of Theorem 3. First, we decompose and bound $|\hat{\rho}^S_{\varphi,n} - \rho^S_{\varphi^*}|$ with

$$|\hat{\rho}^S_{\varphi,n} - \rho^S_{\varphi^*}| \leq |\hat{\rho}^S_{k^*,n} - \rho^S_{k^*}| + |\hat{\rho}^S_{\varphi,n} - \hat{\rho}^S_{\varphi^*}|. \quad (27)$$

In the rest of the proof, we will bound both terms in the right side of (27). Let $G_{k,n}(\kappa) \triangleq \frac{1}{n} \sum_{t=1}^{n} l(x_k, t - \kappa) - G_{\varphi,n}(\kappa) \triangleq \frac{1}{n} \sum_{t=1}^{n} l(x_{\varphi}, t - \kappa)$. We first bound the term $|\hat{\rho}^S_{k^*,n} - \rho^S_{k^*}|$ in (27). Based on Assumption 2, $G_{\varphi,n}(\kappa)$ is monotone decreasing in $\kappa$ and $G_{\varphi,n}(\kappa)$ exists in $[0, M]$ and is upper bounded by $-m_I$. By the differential mean value theorem, there exists an $\xi_n$ between $\hat{\rho}^S_{k^*,n}$ and $\hat{\rho}^S_{\varphi,n}$ that satisfies

$$|G_{\varphi,n}(\hat{\rho}^S_{k^*,n}) - G_{\varphi,n}(\hat{\rho}^S_{\varphi,n})| = |G'_{\varphi,n}(\xi_n)(\hat{\rho}^S_{k^*,n} - \hat{\rho}^S_{\varphi,n})|,$$

so we have

$$|\hat{\rho}^S_{k^*,n} - \rho^S_{k^*}| \leq \frac{1}{m_I} |G_{\varphi,n}(\hat{\rho}^S_{k^*,n}) - G_{\varphi,n}(\hat{\rho}^S_{\varphi,n})|. \quad (28)$$

From Lemma 11, we know that $\hat{\rho}^S_{\varphi,n}$ is the unique solution of $G_{\varphi,n}(\kappa)$, which means that
\[ G_{\varphi, n}(\hat{\rho}_{k^*, n}^S) = 0. \] Restricted to the subset \( \Omega_n^S \), we have

\[
|G_{\varphi, n}(\hat{\rho}_{k^*, n}^S)| = \left| \frac{1}{n} \sum_{t=1}^{n} l(x_{t, t} - \hat{\rho}_{k^*, n}^S) \right|
\]

\[
= \left| \frac{1}{n} \sum_{t=1}^{T_{k^*}(n)} l(x_{k^*, t} - \hat{\rho}_{k^*, n}^S) + \frac{1}{n} \sum_{k \neq k^*} \sum_{t=1}^{T_{k}(n)} l(x_{k, t} - \hat{\rho}_{k^*, n}^S) \right|
\]

\[
\leq \frac{T_{k^*}(n)}{n} G_{k^*, n}^\varphi T_{k^*}(n) (\hat{\rho}_{k^*, n}^S) + \frac{1}{n} \sum_{k \neq k^*} \sum_{t=1}^{T_{k}(n)} l(x_{k, t} - \hat{\rho}_{k^*, n}^S)
\]

\[
\leq \sum_{k \neq k^*} \frac{8M_k^2 M_G}{n \Delta_k} \log \frac{4n^2 K}{\delta},
\]

where the first inequality follows by noting that \( l \) is bounded by \( M_l \). The above inequality implies

\[
|G_{\varphi, n}(\hat{\rho}_{k^*, n}^S) - G_{\varphi, n}(\hat{\rho}_{k^*, n}^S)| \leq \sum_{k \neq k^*} \frac{8M_k^2 M_G}{n \Delta_k} \log \frac{4n^2 K}{\delta},
\]

and so \( |\hat{\rho}_{k^*, n}^S - \hat{\rho}_{k^*, n}^S| \) is bounded by

\[
|\hat{\rho}_{k^*, n}^S - \hat{\rho}_{k^*, n}^S| \leq \sum_{k \neq k^*} \frac{8M_k^2 M_G}{n \Delta_k} \log \frac{4n^2 K}{\delta}.
\]

Meanwhile, restricted to the event \( \Omega_n^S \), the second term in (27) is bounded by

\[
|\hat{\rho}_{k^*, n}^S - \rho_{k^*, n}^S| \leq 2M_l M_G \sqrt{\frac{\log \frac{4n^2 K}{\delta}}{2T_{k^*}(n)}}.
\]

Finally, we have

\[
|\hat{\rho}_{\varphi, n}^S - \rho_{k^*, n}^S| \leq |\hat{\rho}_{k^*, n}^S - \hat{\rho}_{k^*, n}^S| + |\hat{\rho}_{k^*, n}^S - \rho_{k^*, n}^S| \leq \sum_{k \neq k^*} \frac{8M_k^2 M_G}{n \Delta_k} \log \frac{4n^2 K}{\delta} + 2M_l M_G \sqrt{\frac{\log \frac{4n^2 K}{\delta}}{2T_{k^*}(n)}}.
\]

As for the proofs of Theorem 1 and 2, the conclusion follows by noting that both \( \hat{\rho}_{\varphi, n}^S \) and \( \rho_{k^*, n}^S \) lie in \([0, M]\).

\[ \square \]

5 Concluding remarks

In this work, we focus on risk-aware MAB where the objective is a coherent risk measure. We introduce three specific risk measures which are widely investigated in the literature, yet not studied in MAB problems. As our main contribution, we construct an index-based policy for risk-averse MAB and bound its pseudo regret for our three specific risk measures. In particular,
we show that the upper bound on the pseudo regret is of the order of $O(\sqrt{\log n/n})$ which is different from the order $O(\log n/n)$ in risk-neutral case. In our discussion in Section 3.4, we show that this is because the relation between the pseudo regret and the number of pulls of each sub-optimal arm is nonlinear. Meanwhile, we notice that when the risk measures become expectations (by changing the corresponding coefficients), both orders in the risk-averse case and the risk neutral case are consistent. We note that our index policy has a simple form and is thus practical and versatile.

In future work, following the same procedure in this paper, we may extend the scope of our study to incorporate more risk measures. Our scheme in this paper is actually quite general, and only depends on being able to obtain confidence bounds for empirical estimation of risk measures. Moreover, in this work, we need to make specific assumptions on the distribution of each arm as well as on risk measures themselves. For example, for CVaR, we require a continuous CDF. The possibility of relaxing these assumptions is worthy of further study.

Acknowledgement. This research was supported by MOE Tier I grant WBS R266-000-087-112 and MOE Tier I grant WBS R266-000-104-112.

Appendix

Proof of Lemma 3. By Hoeffding’s inequality and noting that $M$ is the uniform upper bound for all the arms, $\forall t = 1, \ldots, n$, we have

$$ P\left\{ \frac{1}{t} \sum_{s=1}^{t} (x_{k,s} - \eta_{k,t}) + - E\left[ (X_k - \eta_{k,t}) \right] \geq M \sqrt{\frac{\log 2n^2K}{2t}} \right\} \leq \frac{\delta}{n^2K}. $$

Thus, by replacing the term $[f_k(\nu_{k0})]^{-1}$ by $m(\alpha)$ in Lemma 2, we have

$$ P\left\{ (\Omega_n^C)^c \right\} \leq \sum_{k=1}^{K} \sum_{t=1}^{n} P\left\{ |\eta_{k,t} - \eta_k| \geq 2m(\alpha) \sqrt{\frac{\log 2n^2K}{2t}} \right\} + \sum_{k=1}^{K} \sum_{t=1}^{n} P\left\{ \frac{1}{t} \sum_{s=1}^{t} (x_{k,s} - \eta_{k,t}) + - E\left[ (X_k - \eta_{k,t}) \right] \geq M \sqrt{\frac{\log 2n^2K}{2t}} \right\} \leq 2\frac{\delta}{n} + \frac{\delta}{n} = 3\frac{\delta}{n}. $$

The conclusion follows.

Proof of Lemma 7. When $x \geq 1$, we have $c \leq ax + bx^{\frac{1}{p}} \leq (a + b)x$, so

$$ x \geq \max \left\{ 1, \frac{c}{a + b} \right\}. $$

When $x \leq 1$, we have $c \leq ax + bx^{\frac{1}{p}} \leq (a + b)x^{\frac{1}{p}}$, so

$$ \min \left\{ 1, \left( \frac{c}{a + b} \right)^{\frac{1}{p}} \right\} \leq x \leq 1. $$
And thus we conclude that \( x \geq \min\{1, \frac{c}{a+b}, (\frac{c}{a+b})^p\} \).

**Proof of Lemma 8.** By Hoeffding’s inequality, and so we have

\[
P \left\{ |\bar{x}_{k,t} - \mu_k| \geq M \sqrt{\frac{\log 4n^2K}{2t}} \right\} \leq \frac{\delta}{2n^2K}.
\]

Note that \(|x_{k,s} - \mu_k|^p\) is bounded by \(M^p\), we have

\[
P \left\{ \left| \frac{1}{t} \sum_{s=1}^t |x_{k,s} - \mu_k|^p - \mathbb{E} |X_k - \mu_k|^p \right| \geq M^p \sqrt{\frac{\log 4n^2K}{2t}} \right\} \leq \frac{\delta}{2n^2K}.
\]

The probability of the event \((\Omega_n^M)^c\) is then bounded above by

\[
P \left\{ (\Omega_n^M)^c \right\} \leq \sum_{k=1}^K \sum_{t=1}^n P \left\{ |\bar{x}_{k,t} - \mu_k| \geq M \sqrt{\frac{\log 4n^2K}{2t}} \right\} + \sum_{k=1}^K \sum_{t=1}^n P \left\{ \left| \frac{1}{t} \sum_{s=1}^t |x_{k,s} - \mu_k|^p - \mathbb{E} |X_k - \mu_k|^p \right| \geq M^p \sqrt{\frac{\log 4n^2K}{2t}} \right\} \leq \frac{\delta}{2n} + \frac{\delta}{2n} = \frac{\delta}{n}.
\]

The conclusion then follows.

**Proof of Lemma 10.** We can directly compute \(G_k(\kappa)\) as

\[
G'_k(\kappa) = \lim_{\Delta \kappa \to 0} \frac{G_k(\kappa + \Delta \kappa) - G_k(\kappa)}{\Delta \kappa} = \lim_{\Delta \kappa \to 0} \int_0^M \left[ l(X_k - \kappa - \Delta \kappa) - l(X_k - \kappa) \right] dF_k(X_k) = \lim_{\Delta \kappa \to 0} \int_0^M \left[ l'(X_k - \kappa) \Delta \kappa + o(\Delta \kappa) \right] dF_k(X_k) = \int_0^M l'(X_k - \kappa) dF_k(X_k).
\]

Based on Assumption 2, \(l'(X_k - \kappa)\) is upper bounded by \(C_l\) and is also lower bounded by noting that \(l(t)\) is strictly increasing in a closed interval \([-M, M]\). So, \(G'_k(\kappa)\) exists in \((0, M)\) and is both upper bounded and lower bounded. The conclusion follows.

**Proof of Lemma 11.** Based on Assumption 2, \(l(x_{k,t} - \kappa)\) is continuous and strictly monotone decreasing in \(\kappa\) for any observation \(x_{k,t}\). As \(X_k\) is bounded in \([0, M]\), noting the monotonicity of \(l(t)\) and \(l(0) = 0\), we have \(l(x_{k,t}) \geq 0\) and \(l(x_{k,t} - M) \leq 0\), so \(\hat{\rho}_{k,n}^{S} \) is the unique solution of

\[
\frac{1}{n} \sum_{t=1}^n l(x_{i,t} - \kappa) = 0,
\]

25
which means \( \hat{\rho}_{k,n}^S \) is an M estimator. Similarly, \( G_k(\kappa) \) is continuous and strictly monotone decreasing because

\[
G_i(\kappa_1) - G_i(\kappa_2) = E[ l(X_i - \kappa_1) - l(X_i - \kappa_2)] > 0
\]

for any \( 0 \leq \kappa_1 < \kappa_2 \leq M \). Let the Lipschitz constant of \( l \) be \( C_l \), then

\[
|G_k(\kappa_1) - G_k(\kappa_2)| = |E[ l(X_k - \kappa_1) - E[l(X_k - \kappa_2)]]|
\]


\[
\leq E[ l(X_k - \kappa_1) - l(X_k - \kappa_2)]
\]

\[
\leq E[C_l|\kappa_1 - \kappa_2|] = C_l|\kappa_1 - \kappa_2|,
\]

so \( \rho_k^S \) is the solution of \( G_k(\kappa) = 0 \). From [DasGupta 2008 Theorem 17.1], we can conclude that \( \hat{\rho}_{k,n}^S \) a.s. \( \rightarrow \rho_k^S \). \( \square \)

Proof of Lemma 13. Based on Lemma 12, we have

\[
P\left\{|\hat{\rho}_{k,t}^S - \rho_k^S| \geq 2M_lM_G\sqrt{\frac{\log \frac{4n^2K}{\delta}}{2t}}\right\} \leq \frac{\delta}{n^2K}.
\]

It follows that

\[
P\{(\Omega_n^S)^c\} \leq \sum_{k=1}^{K} \sum_{t=1}^{n} P\left\{|\hat{\rho}_{k,t}^S - \rho_k^S| \geq 2M_lM_G\sqrt{\frac{\log \frac{4n^2K}{\delta}}{2t}}\right\}
\]

\[
\leq nK \cdot \frac{\delta}{n^2K} = \frac{\delta}{n},
\]

from which we conclude the proof. \( \square \)

References

Acerbi, C. and Tasche, D. (2002), “Expected shortfall: a natural coherent alternative to value at risk,” Economic notes, 31(2), 379–388.

Agrawal, R. (1995), “Sample mean based index policies by o(log n) regret for the multi-armed bandit problem,” Advances in Applied Probability, 27(4), 1054–1078.

Agrawal, R., Hedge, M., and Teneketzis, D. (1988), “Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost,” IEEE Transactions on Automatic Control, 33(10), 899–906.

Arcones, M. A. et al. (1996), “The Bahadur-Kiefer representation for U-quantiles,” The Annals of Statistics, 24(3), 1400–1422.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002), “Finite-time analysis of the multiarmed bandit problem,” Machine learning, 47(2-3), 235–256.
Bahadur, R. R. (1966), “A note on quantiles in large samples,” The Annals of Mathematical Statistics, 37(3), 577–580.

Benati, S. and Rizzi, R. (2007), “A mixed integer linear programming formulation of the optimal mean/value-at-risk portfolio problem,” European Journal of Operational Research, 176(1), 423–434.

Brown, D. B., Giorgi, E. D., and Sim, M. (2012), “Aspirational preferences and their representation by risk measures,” Management Science, 58(11), 2095–2113.

Brown, D. B. and Sim, M. (2009), “Satisficing measures for analysis of risky positions,” Management Science, 55(1), 71–84.

Bubeck, S., Cesa-Bianchi, N., et al. (2012), “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” Foundations and Trends® in Machine Learning, 5(1), 1–122.

DasGupta, A. (2008), Asymptotic theory of statistics and probability, Springer Science & Business Media.

Föllmer, H. and Schied, A. (2002), “Convex measures of risk and trading constraints,” Finance and stochastics, 6(4), 429–447.

Galichet, N., Sebag, M., and Teytaud, O. (2013), “Exploration vs exploitation vs safety: Risk-aware multi-armed bandits,” in Asian Conference on Machine Learning, pp. 245–260.

Jennison, C. and Turnbull, B. W. (1999), Group sequential methods with applications to clinical trials, CRC Press.

Kleinberg, R. D. (2005), “Nearly tight bounds for the continuum-armed bandit problem,” in Advances in Neural Information Processing Systems, pp. 697–704.

Lai, T. L. and Robbins, H. (1985), “Asymptotically efficient adaptive allocation rules,” Advances in applied mathematics, 6(1), 4–22.

Liu, Y., Meskarian, R., and Xu, H. (2017), “Distributionally robust reward-risk ratio optimization with moment constraints,” SIAM Journal on Optimization, 27(2), 957–985.

Maillard, O.-A. (2013), “Robust risk-averse stochastic multi-armed bandits,” in International Conference on Algorithmic Learning Theory, Springer, pp. 218–233.

Mannor, S. and Tsitsiklis, J. N. (2011), “Mean-variance optimization in Markov decision processes,” in Proceedings of the 28th International Conference on International Conference on Machine Learning, Omnipress, pp. 177–184.

Mohri, M. and Munoz, A. (2014), “Optimal regret minimization in posted-price auctions with strategic buyers,” in Advances in Neural Information Processing Systems, pp. 1871–1879.

Robbins, H. (1985), “Some aspects of the sequential design of experiments,” in Herbert Robbins Selected Papers, Springer, pp. 169–177.
Rockafellar, R. T., Uryasev, S., et al. (2000), “Optimization of conditional value-at-risk,” Journal of risk, 2, 21–42.

Ruszczyński, A. and Shapiro, A. (2006), “Optimization of convex risk functions,” Mathematics of operations research, 31(3), 433–452.

Sani, A., Lazaric, A., and Munos, R. (2012), “Risk-aversion in multi-armed bandits,” in Advances in Neural Information Processing Systems, pp. 3275–3283.

Scott, S. L. (2015), “Multi-armed bandit experiments in the online service economy,” Applied Stochastic Models in Business and Industry, 31(1), 37–45.

Shen, Y., Stannat, W., and Obermayer, K. (2013), “Risk-sensitive Markov control processes,” SIAM Journal on Control and Optimization, 51(5), 3652–3672.