Piecewise Stationary Bandits under Risk Criteria

Sujay Bhatt
J.P. Morgan AI Research
New York City, USA
sujaybhatt.hr@gmail.com

Guanhua Fang
School of Management
Fudan University, China
fanggh@fudan.edu.cn

Ping Li
LinkedIn Ads
Bellevue, WA 98004, USA
pinli@linkedin.com

Abstract

Piecewise stationary stochastic multi-armed bandits have been extensively explored in the risk-neutral and sub-Gaussian setting. In this work, we consider a multi-armed bandit framework in which the reward distributions are heavy-tailed and non-stationary, and evaluate the performance of algorithms using general risk criteria. Specifically, we make the following contributions: (i) We first propose a non-parametric change detection algorithm that can detect general distributional changes in heavy-tailed distributions. (ii) We then propose a truncation-based UCB-type bandit algorithm integrating the above regime change detection algorithm to minimize the regret of the non-stationary learning problem. (iii) Finally, we establish the regret bounds for the proposed bandit algorithm by characterizing the statistical properties of the general change detection algorithm, along with a novel regret analysis.

1 INTRODUCTION

Multi-armed Bandits (MAB) are an online sequential decision making framework that effect effort allocation between knowledge exploration and exploitation to identify the optimal choices under uncertainty; see Lattimore and Szepesvári (2020). In the classical version of stochastic bandits, there are multiple 'arms' representing the different unknown reward distributions and an i.i.d reward sequence is associated each of them. A pull of each arm results in observing the associated reward and the learner seeks to find an arm selection strategy that minimizes the *regret*, defined as the difference between the total sum of rewards of an oracle strategy always selecting the arm with largest mean and that of the learner's strategy. The simplicity and generality of this bandit framework has found wide applicability in diverse fields such as internet advertising (Garivier and Moulines, 2011), revenue management (Ferreira et al., 2018), recommender systems (Bouneffouf et al., 2012), clinical trials and medicine (Villar et al., 2015).

There are other real-world applications like healthcare management, finance, etc; see Bouneffouf et al. (2020), where the assumption that the arm distributions do not evolve over time is often violated. This *non-stationary* model was first studied in Kocsis and Szepesvári (2006), where the mean as a function of time evolves in a piecewise stationary manner, and the regret is measured w.r.t the current best arm. Also, in many of the above applications, the data are heavy-tailed (Resnick, 2007; Bhatt et al., 2023), and merely maximizing the expected reward is not always the most desirable objective; some even requiring a consideration of *risk* (Sani et al., 2012).

Motivating Example: Consider the application in healthcare utilization (hospital length of stay or hospital cost – both usually right *skewed*) by scheduling patient category (heart, cancer etc). Frequent selection of a specific action may reduce the expected cost for that action over time, or exogenous factors (ex. covid) may increase it – leading to *non-stationarity*. A bandit framework that accommodates a *risk-sensitive* planner (hospital) seeking to minimize the regret in this non-stationary and possibly heavy-tailed setting, is the object of interest in this paper.

1.1 MAIN CONTRIBUTIONS

We propose a first active adaptation bandit algorithm that has a change detection module which actively monitors general distributional changes, while minimizing the overall risk-sensitive regret. The main contributions are as follows:

1. We propose a first data-driven truncation based non-parametric algorithm to detect general distributional changes in heavy-tailed distributions. Unlike classical sub-Gaussian settings, truncation factors appear in all characteristic parameters of the detection algorithm, namely, the threshold, false alarm probability and de-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

tection delay. This requires a novel statistical analysis of the change detection algorithm.

2. The presence of non-stationarity in the risk-sensitive objective introduces challenges to compare any devised policy against an oracle policy. We propose new sufficient conditions (with examples) that are required for approximating the oracle optimal policy in the non-stationary setting, adding to Cassel et al. (2018).

3. Next, we propose an index based algorithm that actively tracks the distributional changes to identify stationary regimes, and minimizes the *regret* in each of them. The novelty lies in the introduction of a policy-driven truncation method to aid the computation of the index, and subsequent analysis. We characterize[1] $\widetilde{O}(\sqrt{MK/T})$ gap-dependent (normalized [2]) regret bound and $\widetilde{O}((MK)^{\frac{\epsilon}{1+\epsilon}}(T)^{-\frac{\epsilon}{1+\epsilon}})$ worst case regret bound for the proposed algorithm.

4. Experiments illustrate the effectiveness of the proposed adaptive procedures. We report the performance sensitivity w.r.t. various algorithm parameters, where we observe that the proposed algorithm is robust to the shape of the heavy-tailed distributions, percentage of best arm pulls increases with the horizon length, and the distributional changes are detected effectively for a wide range of commonly considered risk measures.

1.2 LITERATURE

A few common objectives considered in risk-sensitive setting include mean-variance (Vakili and Zhao, 2015) and tail-risk measures like VaR (Vakili and Zhao, 2015), CVaR (A. et al., 2020; Tamkin et al., 2019), and general risk measures (Cassel et al., 2018). Bandits with heavy-tailed reward distributions have been the object of interest in various works including Bubeck et al. (2013); Agrawal et al. (2021); Lee et al. (2020). There is an enormous body of literature addressing the subtleties of non-stationary bandit framework; see Cao et al. (2019); Zhou et al. (2021) and the references therein. Notable works in *risk-neutral* non-stationary bandits include passive adaptation Upper Confidence Bound (UCB) algorithms as in Garivier and Moulines (2011) with $O(K\sqrt{MT}\log T)$ and Exp3 algorithms as in Auer et al. (2002) with $O(\sqrt{KMT\log(KT)})$, and active adaptation algorithms as in Liu et al. (2018) obtaining $O(MK\sqrt{T\log T})$ for Exp3 variants and $O(\sqrt{MKT\log T})$ for the UCB variants. All the algorithms mentioned above for piecewise-stationary light-tailed setting, including the one proposed in this paper, require tuning that depends on the number of change points to achieve the state-of-the-art regret $O(\sqrt{MKT\log T})$.

2 PROBLEM FORMULATION

The risk-sensitive non-stationary bandit is described by a $4-$tuple $(\mathcal{K}, \mathcal{T}, U, \{F_{k,t}\}_{k\in\mathcal{K}, t\in\mathcal{T}})$, $\mathcal{K} = \{1, 2, \cdots, K\}$ is a set of K arms, $\mathcal{T} = \{1, 2, \cdots, T\}$ is set of T time steps, U is the risk function, and $F_{k,t}$ is the reward distribution of arm k at time t. The rewards $X_{k,t} \sim F_{k,t}$ are assumed to be independent across arms and time steps. Below we describe the bandit model considered and the objective of the learner.

2.1 BANDIT MODEL

The main features of the bandit model, namely, the piecewise-stationary, heavy-tailed reward distributions, and general risk measures are stated below.

2.1.1 PIECEWISE STATIONARY

Similar to Yu and Mannor (2009), define M to be the number of piecewise-stationary regimes in the reward process:

$$M = 1 + \sum_{t=1}^{T-1} \mathbb{1}\{F_{k,t} \neq F_{k,t+1} \text{ for some } k \in \mathcal{K}\},$$

where $\mathbb{1}$ is the indicator function. Let $\delta_1, \cdots, \delta_{M-1}$ denote the $M-1$ change points with $\delta_0 = 0$ and $\delta_M = T$ for simplicity. For ease of notation, let $\mathcal{M} := \{1, \cdots, M\}$ denote the set of regimes. Note that M is allowed to grow with T and the changes can be asynchronous across arms. We also use the following notation to explicitly identify the distributions: over the i^{th} regime $t \in [\delta_{i-1}+1, \delta_i]$, arm k generates reward $X_k^i \sim F_k^i$.

2.1.2 HEAVY-TAIL REWARDS

We only know that $\mathbb{E}|X_k^i|^{1+\epsilon} := \nu(< \infty)$ for $k \in \mathcal{K}$ in regime $i \in \mathcal{M}$, for some constant $\epsilon \in (0,1]$. The assumption of heavy-tailed rewards necessitates truncation of the empirical distributions to obtain concentration around the underlying mixture distribution. Let the empirical distribution[3] of the reward sequence for arm k and the corresponding random mixture be denoted as $\mathcal{E}_t^k(\cdot) \triangleq \mathcal{E}_t(\{X_{k,i}\}_{i\le t}; \cdot)$ and $\mathcal{F}_t^k \triangleq \frac{1}{t}\sum_{s=1}^{t} F_{k,s}$ respectively. For a sequence of truncation levels $\{b_n(\ge 0)\}$, let the truncated[4] population distribution, $\mathcal{F}_n^{k,\text{trunc}} \triangleq \frac{1}{n}\sum_{s=1}^{n} F_{k,s}^{b_s}$, and the truncated empirical distribution $\mathcal{E}_n^{k,\text{trunc}}(X_{1:n}; \cdot) = \frac{1}{n}\sum_{s=1}^{n} \mathbb{I}_{[\text{Trunc}(X_s, b_s), \infty]}(\cdot)$, where F^b represents the C.D.F. of $\text{Trunc}(X, b)$ with $X \sim F$.

[1] Here M, K, T denote the number of piecewise-stationary regimes, number of arms, and length of the horizon respectively. The notation $\widetilde{O}(\cdot)$ hides the log-factor.

[2] the cumulative regret divided by T

[3] The *empirical distribution* of a real number sequence $x_1, \cdots, x_t$ is the function $\mathcal{E}_t : \mathbb{R}^t \to \mathcal{D}$ such that $\mathcal{E}_t(\{x_i\}_{i\le t}; \cdot) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}_{[x_s, \infty]}(\cdot)$, where $\mathbb{I}_{[a,b]}(\cdot)$ denotes the indicator function on the interval $[a, b]$, i.e, $\mathbb{I}_{[a,b]}(y) = 1$ only if $y \in [a, b]$. Here $\mathcal{D}$ is the space of all distributions on $\mathbb{R}$.

[4] The truncation function of any r.v Y at level $b > 0$ is defined as $\text{Trunc}(Y, b) = \text{sign}(Y)\min\{|Y|, b\}$.

Assumption 1 (Stability)**.** *The following holds for some ψ_1, ψ_2, ψ_3:*
(i) Uniform boundedness, i.e, for $k \in \mathcal{K}$

$$\|\mathcal{F}_n^{k,\mathrm{trunc}} - \mathcal{F}_n^k\| \le \psi_1(n), \textit{ for } n \in \mathcal{Z}^+.$$

(ii) Exponential concentration, i.e, for $k \in \mathcal{K}$

$$\mathbb{P}\Big(\|\mathcal{E}_n^{k,\mathrm{trunc}} - \mathcal{F}_n^{k,\mathrm{trunc}}\| \ge x\Big) \le \exp\Big\{-\frac{nx^2/2}{\psi_2(n) + \psi_3(n)x/3}\Big\}, \textit{ for } n \in \mathcal{Z}^+.$$

Assumption 1 is mild: (i) gives the upper bound on the bias term caused by truncation procedure. (ii) can be viewed as the extended Bernstein inequality. In particular, following Bubeck et al. (2013), we can choose truncation level $b_n = n^{\frac{1}{1+\epsilon}}$. Correspondingly, we can let $\psi_1(n) = c_1 n^{\frac{-\epsilon}{1+\epsilon}}$, $\psi_2(n) = c_2 n^{\frac{1-\epsilon}{1+\epsilon}}$ and $\psi_3(n) = c_3 n^{\frac{1}{1+\epsilon}}$ for some universal constants $c_1 - c_3$. See more details in the supplementary material.

2.1.3 RISK FUNCTION AS PERFORMANCE MEASURE

Classical multi-armed bandit setting considers an additive function of the empirical mean as an optimization criterion. We consider a more general risk function evaluated on the empirical distribution in the optimization problem. This general risk function is viewed as a mapping between seminormed spaces as our analysis relies on certain smoothness properties as in Cassel et al. (2018). Let the space of all distributions on $\mathbb{R}$, denoted as $\mathcal{D}$, be such that $\mathcal{D} \subset \mathcal{L}_{\|\cdot\|}$, a seminormed space associated with the seminorm $\|\cdot\|$.

Definition 1. *A risk function $U : \mathcal{L}_{\|\cdot\|} \to \mathbb{R}$ represents the performance measure over the distributions.*

Let Λ denote the simplex

$$\Lambda \triangleq \Big\{(p_{11}, \cdots, p_{KM}) \in \mathbb{R}^{K\times M} \Big| \sum_{k=1}^{K}\sum_{i=1}^{M} p_{ki} = 1, \; p_{ki} \ge 0, \; \forall\, k \le K, \; i \le M\Big\}.$$

Define the set of all convex combinations of the reward distributions as

$$\mathcal{D}^\Lambda = \Big\{F \in \mathcal{D} : F = \sum_{k=1}^{K}\sum_{i=1}^{M} p_{ij} F_k^i \,\Big|\, p \in \Lambda\Big\}.$$

Assumption 2 (Regularity)**.** *(i) The risk function $U : \mathcal{L}_{\|\cdot\|} \to \mathbb{R}$ is quasi-convex i.e, $U(\alpha F_1 + (1-\alpha)F_2) \le \max\{U(F_1), U(F_2)\}$ for $\alpha \in [0,1]$.*

(ii) There exist $q_1 > 0, q_2 \ge 1$ such that the risk function U admits $\Psi(z) = q_1(z + z^{q_2})$ as a local modulus of continuity for all $F \in \mathcal{D}^\Lambda$,

$$\Big|U(F) - U(G)\Big| \le \Psi\Big(\|F - G\|\Big), \;\; \forall F \in \mathcal{D}^\Lambda, \; G \in \mathcal{L}_{\|\cdot\|}.$$

Quasi-convexity in Assumption 2 is used to define the notion of the 'best' arm for the oracle. The local modulus of continuity property is similar to Cassel et al. (2018), and is required for the decomposition of the regret using the Lipshitz property.

2.2 OBJECTIVE VIA REGRET

For any admissible policy $\boldsymbol{\pi} = \{\pi_t, \; \pi_t \in \mathcal{K}\}$, let $\mathcal{E}_t^{\boldsymbol{\pi}}(\cdot) \triangleq \mathcal{E}_t(\{X_{\pi_i,i}\}_{i\le t}; \cdot)$ and $\mathcal{F}_s^{\boldsymbol{\pi}} \triangleq \frac{1}{t}\sum_{s=1}^{t} F_{\pi_s,s}$. For a given horizon T, the learner aims to minimize the following pseudo-regret

$$\mathcal{R}_{\boldsymbol{\pi}}(T) \triangleq \mathbb{E}\Big[\Big|U(\mathcal{F}_T^{\boldsymbol{\pi}^*}) - U(\mathcal{F}_T^{\boldsymbol{\pi}})\Big|\Big], \; \pi \in \Pi. \qquad (1)$$

The oracle optimal policy $\boldsymbol{\pi}^* = \{\pi^*(i)\}_{i\in\mathcal{M}}$ is defined over the regimes $i \in \mathcal{M}$ such that

$$\pi^*(i) \in \operatorname*{argmax}_{k\in\mathcal{K}} U(F_k^i). \qquad (2)$$

The definition of pseudo-regret boils down to the pseudo-regret defined in the stationary setting ($M = 1$) in Cassel et al. (2018). By the quasi-convexity assumption, we have $U(F_k^i) \le U(F_{\pi^*(i)}^i)$ for all $k \in \mathcal{K}$. So absolute value is vacuous in the stationary setting. However, it is not clear whether the random mixture corresponding the oracle policy so defined also satisfies $U(\mathcal{F}_T^{\boldsymbol{\pi}^*}) \ge U(\mathcal{F}_T^{\boldsymbol{\pi}})$. We find that it is not always the case, and under a stronger sufficient condition of *relational invariance* the relation is indeed satisfied.

Assumption 3. *(Relational Invariance). For any distributions F_1, F_2 and $F_3 \in \mathcal{D}^\Lambda$ and $\alpha \in [0,1]$, it holds*

$$U(\alpha F_1 + (1-\alpha)F_2) \le U(\alpha F_1 + (1-\alpha)F_3) \qquad (3)$$

whenever $U(F_2) \le U(F_3)$.

It is now easy to verify that the random mixture of the oracle policy is such that $U(\mathcal{F}_T^{\boldsymbol{\pi}^*}) \ge U(\mathcal{F}_T^{\boldsymbol{\pi}})$, and the absolute value is not necessary. In particular, if the regime switching can be modeled as a Markov chain, we require the chain to have a unique stationary distribution. In each of the different regimes $\mathcal{M}$, probability $\gamma_\infty(i)$ of regime $i \in \mathcal{M}$ and probability $o(k|i)$ of selecting arm k in regime i. We can see that

$$\max_p U(\sum_k \sum_i F_k^i o(k|i)\gamma_\infty(i))$$

is attained at the desired policy. By quasi-convexity,

$$U(\sum_k F_k^i o(k|i)) \le U(\sum_k F_k^i \mathcal{I}(k = \operatorname*{argmax}_l U(F_l^i))).$$

With $k^*(i) = \operatorname{argmax}_l U(F_l^i)$ and relational invariance,

$$U(\sum_{i,k} F_k^i o(k|i)\gamma_\infty(i)) \le U(\sum_{i,k} F_k^i \mathcal{I}\{k = k^*(i)\}\gamma_\infty(i)).$$

2.3 EXAMPLES OF RISK FUNCTIONS

1. **Linear risk measure**. A general linear risk measure U^{lin} can be written in the following way. There exists a continuous function g such that $U^{\text{lin}}(F) = \int g(x)dF$. Specifically,

$$U^{\text{lin}}(\mathcal{E}_t(\{x_i\}_{i\le t};\cdot)) = \frac{1}{t}\sum_{s=1}^{t} g(x_s)$$

and

$$U^{\text{lin}}(\mathcal{F}_t) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}_{x\sim F_s} g(x),$$

where $\mathcal{F}_t = \frac{1}{t}\sum_{s=1}^{t} F_s$. Then the semi-norm $\|\cdot\| : \mathcal{D} \to \mathbb{R}$ is defined as $\|F\| := |U^{\text{lin}}(F)|$. It is straightforward to check that U^{lin} is stable and relational invariant.

2. **Mean-variance measure**. For a general composite risk measure U^{comp}, it is written as

$$U^{\text{comp}}(F) = h(U^{(1)}(F), \ldots, U^{(L)}(F)).$$

Here function $h : \mathbb{R}^L \to \mathbb{R}$ is a twice-differentiable function and $U^{(1)}, \ldots, U^{(L)}$ are L linear risk measures. The semi-norm under this case is defined as

$$\begin{aligned}\|F\| &:= \|(U^{(1)}(F), \ldots, U^{(L)}(F))\|_2 \\ &= \sqrt{|U^{(1)}(F)|^2 + \ldots + |U^{(L)}(F)|^2}.\end{aligned}$$

For the mean-variance risk, we specifically take $h(x_1, x_2) = x_1 + \rho(x_1^2 - x_2)$ with $x_1 = \int x dF$ and $x_2 = \int x^2 dF$. From Cassel et al. (2018), we know that the mean-variance risk is stable. In addition, if distributions in $\mathcal{D}^{\Lambda}$ have the same mean, then we know that

$$\begin{aligned}&U(\alpha F_1 + (1-\alpha)F_2) - U(\alpha F_1 + (1-\alpha)F_3) \\ &= \int x^2 dF_2 - \int x^2 dF_3 \\ &= U(F_2) - U(F_3).\end{aligned}$$

This implies the relation invariance.

The other common risk measure, CVaR, also satisfies the assumptions. See supplementary for details.

3 NON-PARAMETRIC CHANGE DETECTION FOR HEAVY-TAILS

Typical non-parametric methods for detection of the regime changes are no more sufficient, and new characterizations of truncation dependent thresholds and detection probabilities are needed when considering heavy-tailed rewards. In this section, we propose a general non-parametric detection algorithm whose inputs are truncated and truncation features play an integral role in characterizing statistical properties.

Assumption 4 (Detectability)**.** *Risk function U is detectable, i.e, for any distributions F_k^i and F_k^{i+1}, if $F_k^i \neq F_k^{i+1}$ then $\left|U(F_k^i) - U(F_k^{i+1})\right| > \chi$ for some $\chi > 0$.*

The detectability condition guarantees that the risk of arm k changes by at least χ, when its distribution change at i-th change point. To the best of our knowledge, the current work is the first one to consider a general risk function of heavy-tailed rewards. Therefore we assume Assumption 4 for the ease of analysis. This detectability assumption is a standard assumption in piecewise-stationary bandits (Liu et al., 2018; Cao et al., 2019), wherein the distributional change specified by the risk function (usually the 'mean') is assumed to be significant enough to warrant detection and subsequent classification.

Algorithm 1 Risk-sensitive Heavy-tail Detection (RHD)

```
Given: $2w$—window size, $\beta$—threshold
Rewards $\mathcal{Y}_l \triangleq \{Y_j\}_{j=1}^{w}$ and $\mathcal{Y}_r \triangleq \{Y_j\}_{j=w+1}^{2w}$.
Perform window-driven truncation (Section 3.1) on $\mathcal{Y}_l$
    and $\mathcal{Y}_r$ to get truncated data $\bar{\mathcal{Y}}_l$ and $\bar{\mathcal{Y}}_r$.
procedure RHD($\mathcal{Y}_l \cup \mathcal{Y}_r, \beta$)
if $|U(\mathcal{E}_w(\bar{\mathcal{Y}}_l;\cdot)) - U(\mathcal{E}_w(\bar{\mathcal{Y}}_r;\cdot))| > \beta$ then
    Return True    Change detected
else
    Return False    No change detected
end if
```

3.1 WINDOW-DRIVEN TRUNCATION

Let $2w$ be the window size for change point detection. We hope to use as small number of samples as possible while balancing the delay & false alarm trade-off. For this step, we consider the following truncation method. For each observed reward Y_i $(i = 1, \ldots, 2w)$, we transform it to $\text{Trunc}(Y) = Y\mathbf{1}\{|Y| \le b_w\}$, where b_w is a truncation level only depending on window size w. For example, one possible truncation function we use is given as $b_w \equiv w^{1/(1+\epsilon)}$.

3.2 DISCUSSION & ANALYSIS

In Algorithm 1, if the empirical risk function before the change point is different from the empirical risk function after the change point, then a change is announced. For example, if $U = U^{\text{lin}}$, $\epsilon = 1$, and there is no truncation, then the algorithm is similar to the quickest detection algorithm in Cao et al. (2019). In this case, there is an intuitive trade-off between detection delay and false alarm as a function of β. However, when $\epsilon < 1$, the detection probabilities explicitly rely on the truncation and so does the threshold. Without truncation, it is not possible to derive the concentration without making semi-parametric assumptions.

Let $U(\mathcal{F}_w^{(\cdot),\text{trunc}})$ be the risk of the truncated random mixture of one-half window. We first characterize the gap between $U(\mathcal{F}_w^{l,\text{trunc}})$ and $U(\mathcal{F}_w^{r,\text{trunc}})$. From Assumption 2,

$$\begin{aligned} & |U(\mathcal{F}) - U(\mathcal{F}_w^{(\cdot),\text{trunc}})| \\ \leq\ & \Psi(\|\mathcal{F} - \mathcal{F}_w^{(\cdot),\text{trunc}}\|) \\ =\ & q_1(\|\mathcal{F} - \mathcal{F}_w^{(\cdot),\text{trunc}}\| + \|\mathcal{F} - \mathcal{F}_w^{(\cdot),\text{trunc}}\|^{q_2}) \\ \leq\ & q_1(\psi_1(w) + \psi_1(w)^{q_2}) := \bar{\psi}_1(w) \\ & \text{[It appears in Algorithm 2]} \end{aligned}$$

3.2.1 FALSE ALARM

Again abusing notation, let $\mathcal{E}_w^{k,l} \leftrightarrow \mathcal{E}_w^{k,l,\text{trunc}}$ denote the truncated empirical distribution of the left-half data for arm k in RHD. When there is no change point, we can compute

$$\begin{aligned} & \mathbb{P}(|U(\mathcal{E}_w^{k,l}) - U(\mathcal{E}_w^{k,r})| \geq x) \\ \leq\ & 2\exp\left\{-\frac{w^2(x/4q_1)^2/2}{w\psi_2(w) + \psi_3(w)w(x/4q_1)/3}\right\} \\ & +2\exp\left\{-\frac{w^2(x/4q_1)^{2/q_2}/2}{w\psi_2(w) + \psi_3(w)w(x/4q_1)^{1/q_2}/3}\right\} \\ =:\ & \text{prob}_0(x), \end{aligned} \tag{4}$$

By union bound, the false alarm probability is no greater than $T \cdot \text{prob}_0(\beta)$. In order to ensure $T \cdot \text{prob}_0(\beta)$ to be vanishing as $T \to \infty$, it suffices to have

$$w/\max\{q_1^2\psi_2(w), q_1\psi_3(w)\} \geq C\log T$$

for a large enough constant C.

3.2.2 DETECTION DELAY

When the change happens and we take w such that $q_1(\psi_1(w) + \psi_1(w)^{q_2}) \leq \chi/4$, we know $|U(\mathcal{F}_w^{k,l}) - U(\mathcal{F}_w^{k,r})| \geq \chi/2$. We then can compute

$$\begin{aligned} & \mathbb{P}(|U(\mathcal{E}_w^{k,l}) - U(\mathcal{E}_w^{k,r})| \geq x) \\ \geq\ & 1 - 2\exp\left\{-\frac{w^2((\frac{\chi}{2} - x)/4q_1)^2/2}{w\psi_2(w) + \psi_3(w)w((\frac{\chi}{2} - x)/4q_1)/3}\right\} \\ & -2\exp\left\{-\frac{w^2((\frac{\chi}{2} - x)/4q_1)^{2/q_2}/2}{w\psi_2(w) + \psi_3(w)w((\frac{\chi}{2} - x)/4q_1)^{\frac{1}{q_2}}/3}\right\} \\ :=\ & 1 - \text{prob}_1(x). \end{aligned} \tag{5}$$

In other words, we can detect the change after $\lceil wK/v \rceil$ rounds with probability at least $1 - \text{prob}_1(\beta)$. To ensure that $\text{prob}_1(\beta) \to 0$ when $T \to \infty$, we also require $w/\max\{q_1^2\psi_2(w), q_1\psi_3(w)\} \geq C\log T$.

In summary, the choice of w and β should obey

$$q_1(\psi_1(w) + \psi_1(w)^{q_2}) \leq \chi/4, \tag{6}$$

$$w/\max\{q_1^2\psi_2(w), q_1\psi_3(w)\} \geq C\log T, \tag{7}$$

$$\beta \leq \chi/2. \tag{8}$$

3.3 PROPERTIES OF RHD ALGORITHM

Let the true underlying change points occur at times $\{\delta_1, \delta_2, \cdots\}$, and $\mathcal{V} = \{\nu_1, \nu_2, \cdots\}$ denote change points detected by Algorithm 1.

Proposition 1 (Probability of False Alarm)**.** *Suppose Assumptions 1, 2, and 4 hold. Consider a stationary environment, i.e., the number of regime changes $M = 0$. For a detection threshold β, the false alarm probability $\mathbb{P}(\nu_1 \leq T) \leq KT\text{prob}_0(\beta)$. Here $\text{prob}_0(\beta)$ is defined in Eq. (4).*

Proposition 2 (Expected Detection Delay)**.** *Suppose Assumptions 1, 2, and 4 hold. For an exploration parameter $\zeta \in (0, 1]$ and $L = w\lceil \frac{K}{\zeta} \rceil$, the expected detection delay is $\mathbb{E}\Big[\nu_1 - \delta_1\Big] \leq L(1 - \text{prob}_1(\beta)) + T\text{prob}_1(\beta)$. Here $\text{prob}_1(\beta)$ is defined in Eq. (5).*

4 ALGORITHM FOR REGRET MINIMIZATION

In this section, we propose a confidence based algorithm (Algorithm 2) that actively monitors the regime changes in a non-parametric manner – specified by general distributional changes – and then minimizes the regret in each of the regimes. The algorithm makes use of a novel policy based truncation method to successfully navigate the general setting.

4.1 POLICY-DRIVEN TRUNCATION

Unlike window-driven truncation in RHD where window length $2w$ is set to be a determined number as the input, we need to introduce a data-driven truncation procedure in the index computation step (Line 10) in Algorithm 2 to get a good admissible policy. For this step, we consider the following truncation method. At t-th round, we pull arm π_t and observe reward Y, we transform Y to $\text{Trunc}(Y) = Y\mathbf{1}\{|Y| \leq B_{\pi,t}\}$ where truncation level depends on the admissible policy and the current time index t. For example, suppose at time t, $t_d < t$ is the latest change point time detected by policy π. Then we can set data-dependent truncation level as $B_{\pi,t} = b_{n_{\pi_t}(t)} = (\sum_{s:t\geq s>t_d}\mathbf{1}\{\pi_s = \pi_t\})^{1/(1+\epsilon)}$.

Remark 1 (Note on Truncation)**.** *To handle heavy-tailed data, there are many techniques including "median of means" method (Bubeck et al., 2013), Catoni's M-estimation method (Catoni, 2012; Bhatt et al., 2022d,c), and robust estimation methods (Lee et al., 2020; Bhatt et al., 2022a). However, none of these methods are satisfactory in the current risk-sensitive non-stationary framework. First two methods are relatively hard to compute compared with the truncation method. The first and third methods cannot be easily transformed to a data-riven method. That is, they are relatively less computational friendly in online fashion.*

Lastly, all of three methods suffer the issue that it is hard to quantitatively evaluate the gap between the general risk values *of robustified data and the original one. In other words, the best arm may be changed after implementing those robust techniques. This motivates the truncation methods employed in our framework.*

Algorithm 2 Risk-Sensitive Switching Confidence Bound (RS-SCB)

```
1: $2w-$window size, $\beta-$threshold, T$-$play horizon, $I_t \in \mathcal{K}-$the arm chosen at $t$, $\epsilon \in (0,1]$, $\varrho = 0$.
2: $n_k(t)-$number of arm $k$ pulls up to time $t$, $\zeta \in [0,1]$-exploration factor, $\vartheta > 2-$ input parameter
3: $\mathbb{H}_t = \{X_{I_l,l}\}_{l=1}^t$ and $\mathbb{H}_t^k(\subset \mathbb{H}_t) \triangleq \left\{Y_{k,m}\right\}_{m=1}^{n_k(t)}$.
4: $W_t^k(\subset \mathbb{H}_t^k) \triangleq \left\{Y_{k,m}\right\}_{m=n_k(t)-2w+1}^{n_k(t)}$.
5: for t = 1:T do
6:     $I = (t-\varrho) \mod \lfloor \frac{K}{\zeta} \rfloor$.
7:     if $I \leq K$ then
8:         Pick $I_t = I$.
9:     else
10:        Compute the index, Index$(k,t)$ as (Section 4.1),
           $U\left(\mathcal{E}_{n_k(t)}^{k,\text{trunc}}(\mathbb{H}_t^k;\cdot)\right) + \phi_{n_k(t)}^{-1}\left(\frac{\vartheta \log t}{n_k(t)}\right) + \bar{\psi}_1(n_k(t))$.
11:        Pick $I_t = \text{argmax}_{k\in\mathcal{K}}$ Index$(k,t)$.
12:    end if
13:    Receive reward $X_{I_t,t}$.
14:    Update history $\mathbb{H}_t = \left\{\mathbb{H}_{t-1}, X_{I_t,t}\right\}$.
15:    Update the number of arm pulls $n_{I_t}(t) = n_{I_t}(t) + 1$.
16:    for $k \in \mathcal{K}$ do
17:        Flag = RHD$(W_t^k, \beta)$
18:        if $\big((n_k(t) > 2w)$ & Flag = True$\big)$ then
19:            Redefine history $\mathbb{H}_t^k = \{\}$ for all $k \in \mathcal{K}$.
20:            Reset $\varrho = t$ and $n_k(t) = 0$ for all $k \in \mathcal{K}$.
21:        else
22:            Return
23:        end if
24:    end for
25: end for
```

4.2 DISCUSSION & ANALYSIS

In Algorithm 2, the exploration factor $\zeta \in (0,1]$ determines uniform sampling rate, i.e, it ensures that over a length of M time units, each arm is pulled at least $M/\lceil \frac{K}{\zeta} \rceil$ times, whence the fraction of times **Index** is not computed is approximately ζ. This is to ensure that procedure does not get stuck on an arm just because it has been played most often so far. When the **Index** is updated instead, the best arm is selected in the classical fashion. When the rewards collected exceed $2w$, the regime change detection module checks for changes, and the process repeats. A feature of the active adaptation algorithm is that the influence of the weight of the history is reduced after each 'reset'.

We first explain the function $\phi_{n_k(t)}$ ($\phi_{n_k(t)}^{-1}$) which appears in the line 10 of Algorithm 2. By stability assumption and straightforward calculation, we have the following concentration inequality

$$\mathbb{P}(|U(\mathcal{E}_n^{k,\text{trunc}}) - U(\mathcal{F}_n^{k,\text{trunc}})| \geq x) \leq 2\exp\{-n\phi_n(x)\},$$

where

$$\phi_n(y) = \min\left\{\frac{(y/2q_1)^2}{\psi_2(n) + \psi_3(n)(y/2q_1)/3}, \right. \tag{9}$$

$$\left. \frac{(y/2q_1)^{2/q_2}}{\psi_2(n) + \psi_3(n)(y/2q_1)^{1/q_2}/3}\right\}, \text{ and}$$

$$\phi_n^{-1}(x) = \max\left\{2q_1(\frac{\psi_3(n)x}{3} + \sqrt{\psi_2(n)x}), \right.$$

$$\left. 2q_1\left(\frac{\psi_3(n)x}{3} + \sqrt{\psi_2(n)x\}}\right)^{q_2}\right\}.$$

By the way, it is not hard to see that $\phi_n(\phi_n^{-1}(x)) \geq x$ and $\phi_n(\phi_n^{-1}(x)) \leq 2x$ for any positive x.

4.3 RS-SCB ALGORITHM: PROPERTIES & MAIN RESULTS

In this section, we shall establish the main result, namely, the upper bound on the pseudo regret for Algorithm 2. The algorithm involves two key components: (i) Heavy-tail detection module, and (ii) Arm selection using confidence bounds. Naturally, the pseudo regret has contributions from each of these modules.

Let $\Delta_k := \max_{i\in\mathcal{M}}\left\{U(F_{\pi^*(i)}^i) - U(F_k^i)\right\}$ denote the maximum sub-optimality gap of arm $k \in \mathcal{K}$, the scaling factor $\rho \triangleq \max_{i\in\mathcal{M},k\in\mathcal{K}} \frac{\|F_{\pi^*(i)}^i - F_k^i\|}{\Delta_k}$ denotes the gap-ratio, $G := q_1(1 + D^{q_2-1})$ is a Lipschitz constant with the parameter $D \triangleq \max_{k,l\in\mathcal{K}} \max_{i,j\in\mathcal{M}} \|F_k^i - F_l^j\|$ denoting the diameter of $\mathcal{D}^\Lambda$, and $\eta_k(T) \triangleq \sum_{i=1}^M \sum_{t\in[\delta_{i-1}+1,\delta]} \mathbb{1}\left\{\pi_t = k; k \notin \pi^*(i)\right\}$ denotes the number of sub-optimal plays of arm k over T rounds.

Proposition 3 (Pseudo Regret Decomposition)**.** *The pseudo regret for an arbitrary bandit policy* $\boldsymbol{\pi}$ *is given as*

$$\mathcal{R}_{\boldsymbol{\pi}}(T) \leq \frac{\rho G}{T}\sum_{k=1}^K \Delta_k \mathbb{E}\Big[\eta_k(T)\Big].$$

The pseudo regret is normalized owing to the definition of the mixture distributions. Standard notion of regret is inferred by scaling with T. Let $u_{i,k}$ denote the minimal positive solution to the following inequality,

$$(U(F_{\pi^*(i)}^i) - U(F_k^i))/2 \geq \phi_u^{-1}(\frac{\alpha \log T}{u}) + \bar{\psi}_1(u). \tag{10}$$

Here $u_{i,k}$ can be seen as the length of the phase required to identify the best arm in each regime i.

Theorem 1 (Pseudo Regret RS-SCB)**.** *Suppose Assumptions 1, 2, and 4 hold and choose w, β satisfying (6) - (8). For an exploration parameter $\zeta \in (0, 1]$ in Algorithm 2, the pseudo regret of the RS-SCB algorithm*

$$\mathcal{R}_{\mathrm{RS-SCB}}(T) \leq M\Big(\mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3\Big) + \frac{\rho G \zeta}{K} \sum_{k \in \mathcal{K}} \Delta_k, \quad (11)$$

where the terms $\mathcal{C}_1 \triangleq \sum_{k\in\mathcal{K}} \frac{\rho G}{T}\left(u_k \Delta_k + \Delta_k\right)$, $\mathcal{C}_2 \triangleq \frac{2w\rho G K}{\zeta T} \max_{k\in\mathcal{K}} \Delta_k$, $\mathcal{C}_3 \triangleq \frac{\rho G}{T}\sum_{k\in\mathcal{K}} \Delta_k$, and $u_k = \max_i u_{i,k}$ where $u_{i,k}$'s are defined in Eq. (10).

Term $\mathcal{C}_1$ is the price we pay during the exploration and exploitation phase. $\mathcal{C}_2$ is the price we pay during detection phase. $\mathcal{C}_3$ is the cost induced by false change point detection. The last term in Eq. (11) is the price incurred by the uniform exploration. It is easy to check that the terms in the bound reduce to the terms in Cao et al. (2019) where they only consider sub-Gaussian rewards. Additionally, if we allow exploration factor ζ to depend on M, K, w and T, then we have the following corollaries.

Corollary 1. *Assume Δ_k's are fixed constants and take $\zeta = \sqrt{MwK/T}$, we then have*

$$\mathcal{R}_{\mathrm{RS-SCB}}(T) \leq \frac{\rho G}{T}\sum_k \min\{M u_k, N_k\}\Delta_k + O(\rho G\sqrt{\frac{MwK}{T}})$$

When we take $\psi_1(n) = n^{-\epsilon/(1+\epsilon)}$, $\psi_2(n) = n^{\frac{1-\epsilon}{1+\epsilon}}$ and $\psi_3(n) = n^{\frac{1}{1+\epsilon}}$, we can easily find $w = C(\log T)^{\frac{1+\epsilon}{\epsilon}}$ to satisfy Eq. (7). Thus by ignoring the $\log T$ factor, the bound $O(\rho G\sqrt{\frac{MwK}{T}})$ is nearly optimal since $O(\rho G\sqrt{\frac{MK}{T}})$ is the best instance-dependent lower bound we could do.

Corollary 2. *When Δ_k's are not fixed, we then have the worst case bound*

$$\mathcal{R}_{\mathrm{RS-SCB}}(T) = O\Big(\rho G(\log T)(\frac{MK}{T})^{\epsilon/(1+\epsilon)}\Big)$$

by taking $\zeta = \sqrt{MwK/T}$.

According to the lower bounds established in the supplementary, we know that the proposed RS-SCB algorithm also achieves the optimal instance-independent regret bound if we do not take into account the $\log T$ term. The algorithm structure is similar to Cao et al. (2019). However, unlike the analysis in Cao et al. (2019), we (also) consider weaker assumptions on non-stationarity (see Sec.4.3.1) building on recent results.

Theorem 1 gives the regret bound for a general set of parameters. This is the reason why we do not specify the choice of w and β here. In Corollary 1 and Corollary 2, we provide more specific form. In the setting of Corollary 1, it is assumed that sub-optimality gaps are fixed. Therefore, the regret is dominated by the cost of detecting the change point which is $O(\rho G\sqrt{\frac{MwK}{T}})$. That explains why the regret in this scenario does not explicitly depends on ϵ. In the Corollary 2, Δ_k are no longer assumed to be fixed. Then regret is affected by both exploration via heavy-tailed data and cost of detecting the change point. That is why regret in this case depends on ϵ. For the choice of β, we can simply take it to be $\chi/2$. For the choice of w, we let $w = C(\log T)^{(1+\epsilon)/\epsilon}$ as discussed before.

4.3.1 RELAXATION OF ASSUMPTION 4

In light of recent work in Gur et al. (2014); Besbes et al. (2019); Manegueu et al. (2021), Assumption 4 may be a bit restrictive. Here we generalize the setting and assumption, where the focus is on the severe changes that influence the regret.

Assumption 5. *In addition to Assumption 4, we allow the small fluctuations between change points. That is, there exists a constant $B^* < \chi/4$ such that $|U(F_k^i(t)) - U(F_k^i(t'))| \leq B^*$ for any $t, t' \in [\delta_i, \delta_{i+1})$ and any i.*

Assumption 5 is similar to that in Manegueu et al. (2021) and it gives more flexibility and does not require that distribution $F_k^i(t)$ remains unchanged between two change points δ_i and δ_{i+1}. Under such relaxed setting, we want to design an algorithm only aims to capture those big change points and does not try to detect small changes. By this motivation, we choose w and β satisfy

$$\begin{aligned} q_1(\psi_1(w) + \psi_1(w)^{q_2}) &\leq \chi/8, & (12)\\ w/\max\{q_1^2\psi_2(w), q_1\psi_3(w)\} &\geq C\log T, & (13)\\ \beta &\leq \chi/4, & (14)\end{aligned}$$

and define detection length equation

$$\bar{\Delta}_{i,k}/2 \geq \phi_u^{-1}(\frac{\alpha \log T}{u}) + \bar{\psi}_1(u), \quad (15)$$

where $\bar{\Delta}_{i,k} := \max\{2B^*, \min_{t\in\mathcal{R}_i} U(F^i_{\pi^*(i)}(t)) - \max_{t\in\mathcal{R}_i} U(F_k^i(t))\}$ and $\mathcal{R}_i$ denotes the i-th regime. We further let $\bar{\Delta}_k = \max_i \Delta_{i,k}$ and obtain the following result.

Theorem 2 (Pseudo Regret RS-SCB)**.** *Suppose Assumptions 1, 2, and 5 hold and choose w, β satisfying (12) - (14). For an exploration parameter $\zeta \in (0, 1]$ in Algorithm 2, the pseudo regret of the RS-SCB algorithm*

$$\mathcal{R}_{\mathrm{RS-SCB}}(T) \leq M\Big(\mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3\Big) + \frac{\rho G \zeta}{K} \sum_{k \in \mathcal{K}} \bar{\Delta}_k + 3\rho G B^*. \quad (16)$$

Here the terms $\mathcal{C}_1 \triangleq \sum_{k\in\mathcal{K}} \frac{\rho G}{T}\left(u_k \bar{\Delta}_k + \bar{\Delta}_k\right)$, $\mathcal{C}_2 \triangleq \frac{2w\rho G K}{\zeta T} \max_{k\in\mathcal{K}} \bar{\Delta}_k$ $\mathcal{C}_3 \triangleq \frac{\rho G}{T}\sum_{k\in\mathcal{K}} \bar{\Delta}_k$, and $u_k =$

$\max_i u_{i,k,2}$ *where* $u_{i,k,2}$ *is the minimal positive number satisfying Eq.* (15).

The regret upper bound is looser as a result of this relaxation, and the price for relaxation is approximately $O(\rho G B^*)$ where B^* captures the severity of the changes.

5 EXPERIMENTS

In this section, we provide simulation studies for the proposed RS-SCB algorithm. We would like to emphasize that the main focus of our paper is to integrate three challenging aspects (heavy-tails, non-stationarity, risk) encountered in practical applications (see Motivating Example), and provide a simple algorithm that has good theoretical guarantees. To the best of our knowledge there are no algorithms that provide a fair baseline for heavy tailed piecewise-stationary distributions under risk measures, so we have chosen to illustrate the performance sensitivity w.r.t. various algorithm parameters. For special case of light tail and linear risk measures, the algorithm has similar building blocks as near-optimal non-stationary UCB algorithms like Cao et al. (2019); and we observed that the performance is similar and hence not reported.

Experimental Setting. In all below experiments, we choose truncation level $b_t = \nu^{\frac{1}{1+\epsilon}} t^{\frac{1}{\epsilon}}$, where $\nu = \mathbb{E}[|Z|^{1+\epsilon}], Z \sim F_h$. ($F_h$ is student's-t distribution with degree 2, standard log-normal distribution or Pareto distribution with power index 2.) Consequently, we set $\psi_1(t) = t^{-\epsilon/(1+\epsilon)}$, $\psi_2(t) = t^{(1-\epsilon)/(1+\epsilon)}$ and $\psi_3(t) = t^{1/(1+\epsilon)}$. For a general risk function U, q_1 plays an important role in determining the empirical performance of RS-SCB. Note that small q_1 will lead to many false detections and large q_1 may lead to loose UCB-index which hence slows down the speed to identify the best arm. Hence q_1 can be viewed as an additional tuning parameter for general U in practice.

5.1 CHANGES DETECTABLE UNDER ASSUMPTION 4

We consider the non-stationary bandits with function U being linear risk or CVaR. For each arm k at round t, the reward is specified as $X_{k,t} = \mu_k(t) + \epsilon_t$, where ϵ_t's are independent and follow the same heavy-tailed distribution F_h. In particular, we take F_h to be student's-t distribution, log-normal distribution and pareto distribution. We let $\tilde{M}(:= M-1)$ change points be equally spaced between $[0, T]$. We choose $\mu_k(1) = k - K/2 - 1$ for $k \in [K]$ and set $\chi = 2$ (i.e. $|\mu_k(t+1) - \mu_k(t)| = 2$ for $t \in \{\delta_1, \ldots, \delta_M\}$). Additionally, we fix $K = 5$ and set $T \in \{5, 10, 20, 40, 80\}$ ($\times 10^3$). To apply our method, we choose threshold $\beta = 1$, $w = q_1^2(\log(T))^{\frac{1+\epsilon}{\epsilon}}$, and $\zeta = \sqrt{MKw/T}$. The results are summarized in Figure 1 and Table 1.

From Figure 1, we can see that the percentage of pulling best arm increases as T increases in both settings of linear risk and CVaR. Our method is quite insensitive to the choice of

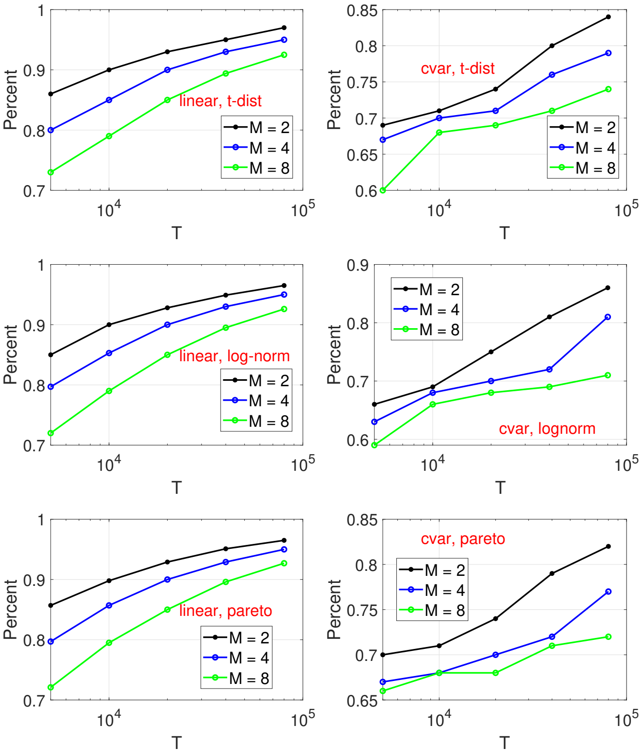

Figure 1: Percentage of pulling best arms. Some hyper parameters are given as: $\epsilon = 1$; $q_1 = 0.5, q_2 = 1$ for linear risk; $\alpha = 0.1$, $q_1 = 2.5, q_2 = 2$ for CVaR.

t-dist	T	5000	10000	20000	40000	80000
linear	$\tilde{M} = 2$	100%	100%	100%	100%	100%
	$\tilde{M} = 4$	99 %	99%	99%	100 %	100%
	$\tilde{M} = 8$	70 %	98%	99%	100%	100%
cvar	$\tilde{M} = 2$	5%	50%	95%	97%	100%
	$\tilde{M} = 4$	1 %	24%	48%	97%	99%
	$\tilde{M} = 8$	1 %	7%	22%	53%	95%
log-norm	T	5000	10000	20000	40000	80000
linear	$\tilde{M} = 2$	98%	100%	100%	100%	100%
	$\tilde{M} = 4$	96%	99%	100%	100 %	100%
	$\tilde{M} = 8$	69%	99%	100%	100%	100%
cvar	$\tilde{M} = 2$	7%	50%	95%	95%	100%
	$\tilde{M} = 4$	0%	21%	49%	95%	96%
	$\tilde{M} = 8$	0%	3%	21%	48%	90%
pareto	T	5000	10000	20000	40000	80000
linear	$\tilde{M} = 2$	98%	100%	100%	100%	100%
	$\tilde{M} = 4$	98%	100%	100%	100%	100%
	$\tilde{M} = 8$	72%	98%	99%	99%	100%
cvar	$\tilde{M} = 2$	12%	45%	95%	100%	100%
	$\tilde{M} = 4$	3%	25%	49%	93%	100%
	$\tilde{M} = 8$	1%	5%	20%	51%	90%

Table 1: Percentage of detection (i.e., the number of times that a change point is detected by RS-SCB divided by the total number of change points) under various cases. Number of underlying change points, $\tilde{M}$, takes values in $\{2, 4, 8\}$.

F_h. That is, RS-SCB is robust for different shapes of heavy-tailed distributions. As number of change points increases, the percentage of pulling best arms decreases as we expect. This is because the exploration factor $\zeta = O(\sqrt{M})$ and the algorithm spends more number of rounds in exploration phase when M gets larger.

From Table 1, we observe that our method performs very well in detecting change points for the linear risk under different choices of $\widetilde{M}$ and T. With T increasing, RS-SCB can also achieve good detection result for CVaR.

5.2 CHANGES DETECTABLE UNDER ASSUMPTION 5

In this section, we provide additional simulation studies for the proposed RS-SCB algorithm under relaxed assumption on the detectable changes.

We still consider the non-stationary bandits with function U being linear risk or CVaR. For each arm k at round t, the reward is specified as $X_{k,t} = \mu_k(t) + \omega(t) + \epsilon_t$, where ϵ_t's are independent and follow the same heavy-tailed distribution F_h. Since our method is quite robust to the choice of F_h, we here only take F_h to be student's-t distribution for illustrative purpose. We again let $\widetilde{M}(:= M-1)$ change points be equally spaced between $[0, T]$. We choose $\mu_k(1) = k - K/2 - 1$ for $k \in [K]$ and set $\chi = 2$ (i.e. $|\mu_k(t+1) - \mu_k(t)| = 2$ for $t \in \{\delta_1, \ldots, \delta_M\}$). We further select $\omega(t) = B^* \sin(2\pi \widetilde{M} t/T)$. The perturbation level B^* takes value from $\{0.2, 0.5, 1.0\}$. As before, we still fix $K = 5$ and set $T \in \{5, 10, 20, 40, 80\}$ $(\times 10^3)$. To apply our method, we choose threshold $\beta = 1$, $w = q_1^2(\log(T))^{\frac{1+\epsilon}{\epsilon}}$, and $\zeta = \sqrt{MKw/T}$. Results are summarized in Figure 2. (Remark: optimal tuning of w, ζ requires the knowledge of T. When T is unknown, we can split into geometrically spaced time intervals $[2^k, 2^{k+1} - 1]$ $(k \geq 0)$ and use different w, ζ in each interval.)

From Figure 2, we can see that the percentage of pulling best arm increases as T increases in both settings of linear risk and CVaR. Our method is also quite insensitive to the choice of B^*. The performance does not degrade too much when perturbation level, B^*, increases from 0.2 to 1.0.

6 CONCLUSION & FUTURE WORK

We provided an index based algorithm that is truncation-based and non-parametric, and also enjoys nearly-optimal regret bound in both gap-dependent/independent sense. The algorithm combines an actively adapting non-parametric change point detection algorithm that is designed to identify general distributional changes in heavy-tailed distributions along with novel data-driven truncation to provide a tight characterization. We also characterized the delay and false alarm probabilities (as a function of truncation) for heavy-tailed distributions that is of independent interest.

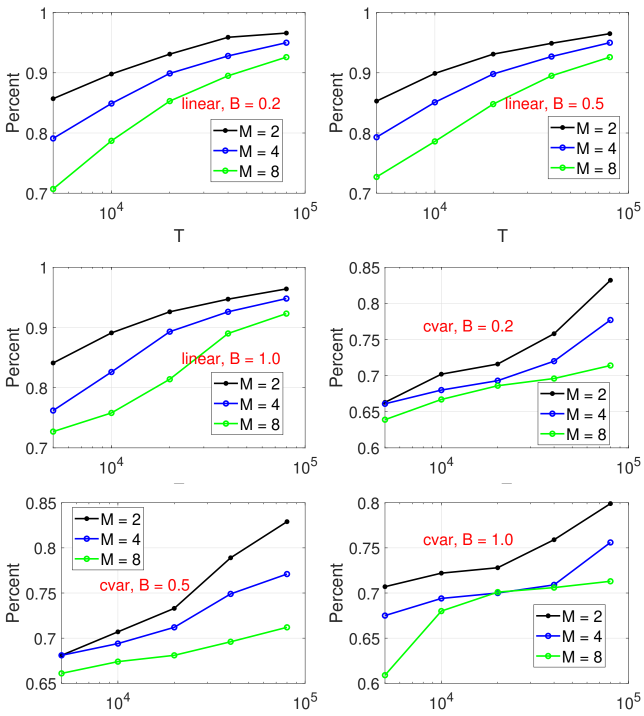

Figure 2: Percentage of pulling best arms for generalized settings. ("1" means 100%.) Hyper parameters are the same as before: $\epsilon = 1$; $q_1 = 0.5, q_2 = 1$ for linear risk; $\alpha = 0.1, q_1 = 2.5, q_2 = 2$ for CVaR.

Below we comment on the limitations of this work and these naturally lead to future research directions. (i) We implicitly assume the knowledge of M, the number piecewise-stationary regimes as in previous works. Recent works, for example (Auer et al., 2019; Chen et al., 2019; Besson et al., 2022), relax this assumption (in linear risk setting) while still obtaining similar regret guarantees. Exploring this direction for the general framework is a worthwhile endeavor. (ii) An important trade-off is that while the active adaptation algorithm provides a $\widetilde{O}(\sqrt{K})$ regret, it requires strong assumptions on the risk-measures. A simple sliding-window type approach that leads to a slightly higher regret $\widetilde{O}(K)$, but works under more general conditions might be suitable in real-world applications. (iii) We are only able to establish a $\widetilde{\Omega}(\frac{\sqrt{MK}}{\sqrt{T}})$ lower bound under an additional assumption on the risk function, namely, the U-structure. It leaves an open question: is the gap-dependent lower bound order for any type of risk (ex. CVaR)? (iv) We assume the knowledge of the moment bound parameter ϵ and the horizon T in the algorithm parameters tuning and implementation. Robustification w.r.t these parameters and using anytime confidence bounds as in Bhatt et al. (2022b) might be worth considering for future work. (v) Adopting techniques in the current paper into other bandit settings (e.g. adversarial bandit (Bubeck and Slivkins, 2012), contextual bandit (Li et al., 2010), fairness bandit (Fang et al., 2022), etc.) might also be an interesting research problem.

References

Prashanth L. A., Krishna P. Jagannathan, and Ravi Kumar Kolla. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5577–5586, Virtual Event, 2020.

Shubhada Agrawal, Sandeep Juneja, and Wouter M. Koolen. Regret minimization in heavy-tailed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 26–62, Boulder, CO, 2021.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Peter Auer, Yifang Chen, Pratik Gajane, Chung-Wei Lee, Haipeng Luo, Ronald Ortner, and Chen-Yu Wei. Achieving optimal dynamic regret for non-stationary bandits without prior information. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 159–163, Phoenix, AZ, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4): 319–337, 2019.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *J. Mach. Learn. Res.*, 23(77):1–40, 2022.

Sujay Bhatt, Guanhua Fang, and Ping Li. Offline change detection under contamination. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 191–201, Eindhoven, The Netherlands, 2022a.

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Catoni-style confidence sequences under infinite variance. *arXiv preprint arXiv:2208.03185*, 2022b.

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Minimax m-estimation under adversarial contamination. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1906–1924, Baltimore, MD, 2022c.

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Nearly optimal catoni's m-estimator for infinite variance. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1925–1944, Baltimore, MD, 2022d.

Sujay Bhatt, Ping Li, and Gennady Samorodnitsky. Extreme bandits using robust statistics. *IEEE Trans. Inf. Theory*, 69(3):1761–1776, 2023.

Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *Proceedings of the 19th International Conference on Neural Information Processing (ICONIP), Part III*, pages 324–331, Doha, Qatar, 2012.

Djallel Bouneffouf, Irina Rish, and Charu C. Aggarwal. Survey on applications of multi-armed and contextual bandits. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, Glasgow, United Kingdom, 2020.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 42.1–42.23, Edinburgh, Scotland, 2012.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Trans. Inf. Theory*, 59(11): 7711–7717, 2013.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 418–427, Naha, Okinawa, Japan, 2019.

Asaf B. Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Proceedings of the Conference On Learning Theory (COLT)*, pages 1295–1306, Stockholm, Sweden, 2018.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 696–726, Phoenix, AZ, 2019.

Guanhua Fang, Ping Li, and Gennady Samorodnitsky. On penalization in stochastic multi-armed bandits. *arXiv preprint arXiv:2211.08311*, 2022.

Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Oper. Res.*, 66(6):1586–1602, 2018.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT)*, pages 174–188, Espoo, Finland, 2011.

Yonatan Gur, Assaf Zeevi, and Omar Besbes. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems (NIPS)*, pages 199–207, Montreal, Canada, 2014.

Levente Kocsis and Csaba Szepesvári. Discounted ucb. In *Proceedings of the 2nd PASCAL Challenges Workshop*, volume 2, pages 51–134, 2006.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 661–670, Raleigh, NC, 2010.

Fang Liu, Joohyun Lee, and Ness B. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 3651–3658, New Orleans, LA, 2018.

Anne Gael Manegueu, Alexandra Carpentier, and Yi Yu. Generalized non-stationary bandits. *arXiv preprint arXiv:2102.00725*, 2021.

Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

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

Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *Proceedings of 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335, Monticello, IL, 2015.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical S'cience: a Review Journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1177–1184, Montreal, Canada, 2009.

Xiang Zhou, Yi Xiong, Ningyuan Chen, and Xuefeng Gao. Regime switching bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4542–4554, virtual, 2021.

Supplementary Materials

Notation

For reader convenience, we collect all important notion in this appendix. At time t, the learner/ bandit algorithm chooses action π_t and receives reward $X_{\pi_t,t}$. An admissible bandit policy $\boldsymbol{\pi} = \{\pi_t\}$ is a random process recursively defined as

$$\pi_t \triangleq \pi_t\Big(\pi_1, \pi_2, \cdots, \pi_{t-1}, X_{\pi_1,1}, X_{\pi_2,2}, \cdots, X_{\pi_{t-1},t-1}\Big),$$

where the policy $\pi_t : \mathcal{H}_t \to \mathcal{K}$ is a function of the past history of actions and observations. Formally, the bandit policy is measurable w.r.t to the sigma-algebra

$$\mathcal{H}_t \triangleq \Sigma\Big(\pi_1, \pi_2, \cdots, \pi_{t-1}, X_{\pi_1,1}, X_{\pi_2,2}, \cdots, X_{\pi_{t-1},t-1}\Big).$$

Let $\mathcal{D}$ be the space of all distributions on $\mathbb{R}$. The *empirical distribution* of a real number sequence $x_1, \cdots, x_t$ is the function $\mathcal{E}_t : \mathbb{R}^t \to \mathcal{D}$ such that

$$\mathcal{E}_t(\{x_i\}_{i\le t}; \cdot) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}_{[x_s,\infty]}(\cdot),$$

where $\mathbb{I}_{[a,b]}(\cdot)$ denotes the indicator function on the interval $[a, b]$. The empirical distribution of the reward sequence under policy $\boldsymbol{\pi}$, and arm k are defined as

$$\mathcal{E}_t^{\boldsymbol{\pi}}(\cdot) \triangleq \mathcal{E}_t(\{X_{\pi_i,i}\}_{i\le t}; \cdot) \text{ and } \mathcal{E}_t^k(\cdot) \triangleq \mathcal{E}_t(\{X_{k,i}\}_{i\le t}; \cdot)$$

respectively. Let the corresponding *random mixture* distributions for policy $\boldsymbol{\pi}$ and $k \in \mathcal{K}$ be given as

$$\mathcal{F}_s^{\boldsymbol{\pi}} \triangleq \frac{1}{t}\sum_{s=1}^{t} F_{\pi_s,s} \text{ and } \mathcal{F}_t^k \triangleq \frac{1}{t}\sum_{s=1}^{t} F_{k,s}$$

respectively. Define the truncation function at level $b > 0$ as

$$\text{Trunc}(Y, b) = \text{sign}(Y)\min\{|Y|, b\}.$$

Let F^b represent the C.D.F. of $\text{Trunc}(Y, b)$ with $Y \sim F$, and $(b_n; n = 1, \ldots)$ denote a sequence of truncation levels. For arm k, we define the truncated population distribution,

$$\mathcal{F}_n^{k,\text{trunc}} \triangleq \frac{1}{n}\sum_{s=1}^{n} F_{k,s}^{b_s},$$

and, for any $n \in \mathcal{Z}^+$, the truncated empirical distribution

$$\mathcal{E}_n^{k,\text{trunc}}(X_{1:n}; \cdot) = \frac{1}{n}\sum_{s=1}^{n} \mathbb{I}_{[\text{Trunc}(X_s,b_s),\infty]}(\cdot).$$

Let Π denote the set of all admissible policies. Given any policy $\pi \in \Pi$, we can also define the policy-dependent truncated population distribution as

$$\mathcal{F}_n^{\boldsymbol{\pi},\text{trunc}} \triangleq \frac{1}{n}\sum_{s=1}^{n} F_{\pi_s,s}^{B_{\pi,s}}$$

and the policy-dependent truncated empirical distribution as

$$\mathcal{E}_n^{\boldsymbol{\pi},\text{trunc}} = \frac{1}{n}\sum_{s=1}^{n} \mathbb{I}_{[\text{Trunc}(X_s,B_{\boldsymbol{\pi},s}),\infty]}(\cdot).$$

Here $B_{\pi,s}$ takes value in $(b_n; n = 1, \ldots)$. It depends on time index s and policy π and thus is history-dependent.

Let the space of all distributions on $\mathbb{R}$ denoted as $\mathcal{D}$ be such that $\mathcal{D} \subset \mathcal{L}_{\|\cdot\|}$, a seminormed space associated with the seminorm $\|\cdot\|$. Let Λ denote the simplex in $\mathbb{R}^{K\times M}$:

$$\Lambda \triangleq \Big\{(p_{11}, \cdots, p_{KM}) \;\in\; \mathbb{R}^{K\times M} \Big| \sum_{k=1}^{K}\sum_{i=1}^{M} p_{ki} = 1,\; p_{ki} \geq 0,\; \forall\, k \leq K,\; i \leq M\Big\}.$$

Define the set of all convex combinations of the reward distributions as

$$\mathcal{D}^{\Lambda} = \Big\{F \in \mathcal{D} : F = \sum_{k=1}^{K}\sum_{i=1}^{M} p_{ij} F_k^i \;\Big|\; p \in \Lambda\Big\}.$$

The following parameters associated with *regret*. Let

$$\Delta_k := \max_{i\in\mathcal{M}} \Big\{U(F^i_{\pi^*(i)}) - U(F^i_k)\Big\}$$

denote the maximum sub-optimality gap of arm $k \in \mathcal{K}$, and

$$\rho \triangleq \max_{i\in\mathcal{M}, k\in\mathcal{K}} \frac{\|F^i_{\pi^*(i)} - F^i_k\|}{\Delta_k}$$

denote the gap-ratio, $D \triangleq \max_{k,l\in\mathcal{K}} \max_{i,j\in\mathcal{M}} \|F_k^i - F_l^j\|$ denote the diameter of $\mathcal{D}^{\Lambda}$,

$$G := q_1(1 + D^{q_2 - 1})$$

is a Lipschitz constant and

$$\eta_k(T) \triangleq \sum_{i=1}^{M} \sum_{t\in[\delta_{i-1}+1, \delta]} \mathbb{1}\Big\{\pi_t = k; k \notin \pi^*(i)\Big\}$$

denote number of sub-optimal plays of arm k over T rounds.

In this work, we always view gap-ratio ρ and diameter G as two absolute constants. For example, in linear risk case with bounded reward between $[0, 1]$, then $\rho \equiv 1, G \equiv 2$.

7 Comments on Assumption 1

Here we explain the reasons why Assumption 1 is not restrictive. For (i) uniform boundedness, in the linear risk settings, we have

$$\begin{aligned} \|\mathcal{F}_n^{k,\mathrm{trunc}} - \mathcal{F}_n^k\| &= \frac{1}{n}\Big|\sum_{t=1}^{n} \mathbb{E}[\mathrm{Trunc}(X_t, b_t) - X_t]\Big| \\ &\leq \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[|X_t|\mathbf{1}_{|X_t|\geq b_t}] \\ &\leq \frac{1}{n}\sum_{t=1}^{n} \frac{\nu}{b_t^{\epsilon}} \leq 2\nu n^{-\frac{\epsilon}{1+\epsilon}}, \end{aligned} \tag{17}$$

where we choose $b_t = t^{1/(1+\epsilon)}$ for $t = 1, \dots, n$. Therefore, we can take $\psi_1(n) = 2\nu n^{-\frac{\epsilon}{1+\epsilon}}$. Similarly, in the conditional value at risk (CVaR) setting, we have

$$\begin{aligned}
& \|\mathcal{F}_n^{k,\text{trunc}} - \mathcal{F}_n^k\| \\
= \ & \max\{\|\frac{1}{n}\sum_{t=1}^n (F_t^{b_t} - F_t)\|_\infty, |\frac{1}{n}\sum_t \mathbb{E}[X_{t+} - \text{Trunc}(X_t, b_t)_+]|, |\frac{1}{n}\sum_t \mathbb{E}[X_{t-} - \text{Trunc}(X_t, b_t)_-]|\} \\
\le \ & \max\{\frac{1}{n}\sum_{t=1}^n \|F_t^{b_t} - F_t\|_\infty, 2\nu n^{-\epsilon/(1+\epsilon)}\} \\
\le \ & \max\{\frac{1}{n}\sum_{t=1}^n \min\{1, \frac{\nu}{|t-\nu|}\}, 2\nu n^{-\epsilon/(1+\epsilon)}\} \\
\le \ & \max\{\frac{\nu(\log n + 2)}{n}, 2\nu n^{-\epsilon/(1+\epsilon)}\} \ \ (\text{since} \sum_{i=1}^n \frac{1}{i} \le \log n + 1), \qquad (18)
\end{aligned}$$

where we still choose $b_t = t^{1/(1+\epsilon)}$. Thus we can take $\psi_1(n) = \max\{\frac{\nu(\log n+2)}{n}, 2\nu n^{-\epsilon/(1+\epsilon)}\}$.

For (ii) exponential concentration in Assumption 1, we can apply Bernstein inequality to the following bounded independent random variables, $\text{Trunc}(X_t, b_t) - \mathbb{E}[\text{Trunc}(X_t, b_t)]$'s ($t = 1, \dots, n$) (or $\text{Trunc}(X_t, b_t)_+ - \mathbb{E}[\text{Trunc}(X_t, b_t)_+]$, $\text{Trunc}(X_t, b_t)_- - \mathbb{E}[\text{Trunc}(X_t, b_t)_-]$), and apply Dvoretzky–Kiefer–Wolfowitz (DKW) inequality to $\frac{1}{n}\sum_{t=1}^n (\mathbb{I}[\text{Trunc}(X_t, b_t), \infty](\cdot) - F_t^{b_t}(\cdot))$. Hence

$$\begin{aligned}
& \mathbb{P}(\|\mathcal{E}_n^{k,trunc} - \mathcal{F}_n^{k,trunc}\| \ge x) \\
\le \ & 2\exp\{-\frac{n^2x^2/2}{\nu\sum_t b_t^{1-\epsilon} + n\max_t b_t x/3}\} + 2\exp\{-2nx^2\} \qquad (19) \\
\le \ & 4\exp\{-\frac{nx^2/2}{\nu n^{\frac{1-\epsilon}{1+\epsilon}} + n^{\frac{1}{1+\epsilon}} x/3}\}, \qquad (20)
\end{aligned}$$

where Eq. (19) uses the fact that $\mathbb{E}[|\text{Trunc}(X_t, b_t)|^2] \le \mathbb{E}[|\text{Trunc}(X_t, b_t)|^{1-\epsilon}|\text{Trunc}(X_t, b_t)|^{1+\epsilon}] \le b_t^{1-\epsilon}\nu$, Eq. (20) chooses $b_t = t^{1/(1+\epsilon)}$. Hence, we can choose $\psi_2(n) = c_2 n^{\frac{1-\epsilon}{1+\epsilon}}$ and $\psi_3(n) = c_3 n^{\frac{1}{1+\epsilon}}$. This explains why Assumption 1 is mild.

8 Other examples satisfying required assumptions

1. **Conditional Value at Risk (CVaR)**. CVaR_α is the average reward below percentile level α and its formula is given by

$$U^{\text{CVaR}_\alpha}(F) = \max_{z\in\mathbb{R}} z - \frac{1}{\alpha}\int_{-\infty}^{\alpha} F(x)dx. \qquad (21)$$

Then the semi-norm is defined as

$$\|F\| = \max\{\|F\|_\infty, |\int_{-\infty}^0 xdF(x)|, |\int_0^\infty xdF(x)|\}. \qquad (22)$$

The stability of CVaR holds by applying DKW or Bretagnolle inequality for $\|\mathcal{E}_t^k - \mathcal{F}_t^k\|_\infty$ and Bernstein's inequality for $|\int_{-\infty}^0 xd(\mathcal{E}_t^k - \mathcal{F}_t^k)|, |\int_0^\infty xd(\mathcal{E}_t^k - \mathcal{F}_t^k)|$.

When $M = 2$ with $U(F^{i_1}) < U(F^{i_2})$, we know that any distribution $F \in \mathcal{D}^\Lambda$ can be written as $\alpha_F F^{i_1} + (1-\alpha_F)F^{i_2}$ with $0 \le \alpha_F \le 1$. Thus $U(\alpha F_1 + (1-\alpha)F_2) \le U(\alpha F_1 + (1-\alpha)F_3)$ is equivalent to $\alpha(1-\alpha_{F_1}) + (1-\alpha)(1-\alpha_{F_2}) < \alpha(1-\alpha_{F_1}) + (1-\alpha)(1-\alpha_{F_3})$. It is further equivalent to $(1-\alpha_{F_2}) < (1-\alpha_{F_3})$, i.e., $U(F_2) < U(F_3)$. This leads to the relational invariance.

At the end of this section, we show that CVaR satisfies the regularity assumption for completeness. First by Lemma 9 in Cassel et al. (2018), we have that

$$|U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)| \le \frac{2\|F\| + \|F-G\|}{\min\{\alpha, 1-\alpha\}}, \qquad (23)$$

and

$$\begin{aligned} 0 &\leq U^{CVaR_\alpha}(F) - U^{CVaR_\alpha}(G) + \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} (G(x) - F(x))dx \\ &\leq \frac{1}{\alpha}\|F - G\||U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)|. \end{aligned} \tag{24}$$

Also note that

$$\begin{aligned} |\frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} (G(x) - F(x))dx| &\leq |\frac{1}{\alpha}\int_{-\infty}^{0} (G(x) - F(x))dx| + |\frac{1}{\alpha}\int_{0}^{U_\alpha^{VaR}(F)} (G(x) - F(x))dx| \\ &\leq \frac{1}{\alpha}(\|F - G\| + |U^{VaR_\alpha}(F)|\|F - G\|) \\ &= \frac{\|F - G\|}{\alpha}(1 + |U^{VaR_\alpha}(F)|). \end{aligned} \tag{25}$$

Combined with Eq. (24), we have

$$|U^{CVaR_\alpha}(F) - U^{CVaR_\alpha}(G)| \leq \frac{\|F - G\|}{\alpha}(1 + |U^{VaR_\alpha}(F)| + |U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)|). \tag{26}$$

Further by Eq. (23), we get that

$$\begin{aligned} |U^{CVaR_\alpha}(F) - U^{CVaR_\alpha}(G)| &\leq \frac{\|F - G\|}{\alpha}(1 + |U^{VaR_\alpha}(F)| + \frac{2\|F\| + \|F - G\|}{\min\{\alpha, 1 - \alpha\}}) \\ &\leq \frac{1}{\alpha}(1 + |U^{VaR_\alpha}(F)| + \frac{\max 1, 2\|F\|}{\min\{\alpha, 1 - \alpha\}})(\|F - G\| + \|F - G\|^2) \\ &\leq (4 + Q)/(\alpha \min\{\alpha, 1 - \alpha\})(\|F - G\| + \|F - G\|^2). \end{aligned} \tag{27}$$

Thus, we can take $q_1 = (4 + Q)/(\alpha \min\{\alpha, 1 - \alpha\})$ and $q_2 = 2$ when we consider the family of distribution with α-quantile bounded by Q.

9 Missing Details in the Main Context

Example 1. For linear risk function, we obtain that $\psi_1(w) = \nu w^{-\epsilon/(1+\epsilon)}$ by the following fact, as $\mathbb{E}[|X|^{1+\epsilon}] \leq \nu$,

$$\|\mathcal{F} - \mathcal{F}_w^{\text{trunc}}\| = |\mathbb{E}[X - \psi(X)]| = |\mathbb{E}[X\mathbf{1}\{|X| > w^{1/(1+\epsilon)}\}]| \leq \nu w^{-\epsilon/(1+\epsilon)}, \tag{28}$$

Example 2. For CVaR risk function, we obtain that $\psi_1(w) = \max\{\frac{\nu}{|w-\nu|}, \nu w^{-\epsilon/(1+\epsilon)}\}$ by the following,

$$\begin{aligned} \|\mathcal{F} - \mathcal{F}_w^{\text{trunc}}\| &= \max\{\|\mathcal{F} - \mathcal{F}_w^{\text{trunc}}\|_\infty, |\mathbb{E}[X_+ - \psi(X_+)]|, |\mathbb{E}[X_+ - \psi(X_+)]|\} \\ &\leq \max\{|1 - \frac{1}{1 - \mathbb{P}(|X| \geq w^{1/(1+\epsilon)})}|, \nu w^{-\epsilon/(1+\epsilon)}\} \leq \max\{|1 - \frac{1}{1 - \nu/w}|, \nu w^{-\epsilon/(1+\epsilon)}\}, \end{aligned}$$

where X_+ and X_- are positive and negative parts of X, respectively. Therefore, we arrive at

$$|U(\mathcal{F}_l^{w,\text{trunc}}) - U(\mathcal{F}_r^{w,\text{trunc}})| \geq |U(\mathcal{F}_w^l) - U(\mathcal{F}_w^r)| - 2\bar{\psi}_1(w). \tag{29}$$

In other words, $|U(\mathcal{F}_w^{l,\text{trunc}}) - U(\mathcal{F}_w^{r,\text{trunc}})|$ has gap at least $\chi - 2\bar{\psi}_1(w)$ when $|U(\mathcal{F}_w^l) - U(\mathcal{F}_w^r)| = \chi$. When w is large enough, we have $|U(\mathcal{F}_w^{l,\text{trunc}}) - U(\mathcal{F}_w^{r,\text{trunc}})| \geq \frac{1}{2}|U(\mathcal{F}_w^l) - U(\mathcal{F}_w^r)| = \frac{1}{2}\chi$.

The false alarm probability Eq. 8 is calculated by

$$\begin{aligned} \mathbb{P}(|U(\mathcal{E}_w^{k,l}) - U(\mathcal{E}_w^{k,r})| \geq x) &\leq \mathbb{P}(|U(\mathcal{E}_w^{k,l}) - U(\mathcal{F}_w^{k,l})| \geq x/2) + \mathbb{P}(|U(\mathcal{E}_w^{k,r}) - U(\mathcal{F}_w^{k,r})| \geq x/2) \\ &\leq 2\mathbb{P}(\|\mathcal{E}_w^{k,l} - \mathcal{F}_w^{k,l}\| \geq \frac{x}{4q_1}) + 2\mathbb{P}(\|\mathcal{E}_w^{k,r} - \mathcal{F}_w^{k,r}\| \geq (\frac{x}{4q_1})^{1/q_2}) \\ &\leq 2\exp\Big\{-\frac{w^2(x/4q_1)^2/2}{w\psi_2(w) + \psi_3(w)w(x/4q_1)/3}\Big\} \\ &\quad + 2\exp\Big\{-\frac{w^2(x/4q_1)^{2/q_2}/2}{w\psi_2(w) + \psi_3(w)w(x/4q_1)^{1/q_2}/3}\Big\} \\ &:= \text{prob}_0(x). \end{aligned} \tag{30}$$

The detection power Eq. 9 is given by the following calculation,

$$\begin{aligned}
& \mathbb{P}(|U(\mathcal{E}_w^{k,l}) - U(\mathcal{E}_w^{k,r})| \geq x) \\
\geq \quad & 1 - \mathbb{P}(|U(\mathcal{E}_w^{k,l}) - U(\mathcal{F}_w^{k,l})| \geq (\chi/2 - x)/2) - \mathbb{P}(|U(\mathcal{E}_w^{k,r}) - U(\mathcal{F}_w^{k,r})| \geq (\chi/2 - x)/2) \\
\geq \quad & 1 - 2\Big\{\mathbb{P}(\|\mathcal{E}_w^{k,l} - \mathcal{F}_w^{k,l}\| \geq \frac{(\chi/2 - x)}{4q_1}) + \mathbb{P}(\|\mathcal{E}_w^{k,l} - \mathcal{F}_w^{k,l}\| \geq (\frac{(\chi/2 - x)}{4q_1})^{1/q_2})\Big\} \\
\geq \quad & 1 - 2\exp\Big\{-\frac{w^2((\frac{\chi}{2} - x)/4q_1)^2/2}{w\psi_2(w) + \psi_3(w)w((\frac{\chi}{2} - x)/4q_1)/3}\Big\} \\
& 2\exp\Big\{-\frac{w^2((\frac{\chi}{2} - x)/4q_1)^{2/q_2}/2}{w\psi_2(w) + \psi_3(w)w((\frac{\chi}{2} - x)/4q_1)^{\frac{1}{q_2}}/3}\Big\} \\
:= \quad & 1 - \text{prob}_1(x). \qquad (31)
\end{aligned}$$

Derivation of $\phi_n(x)$ is given by

$$\begin{aligned}
& \mathbb{P}(|U(\mathcal{E}_n^{k,\text{trunc}}) - U(\mathcal{F}_n^{k,\text{trunc}})| \geq x) \\
\leq \quad & \mathbb{P}(\|\mathcal{E}_n^{k,\text{trunc}} - \mathcal{F}_n^{k,\text{trunc}}\| \geq \frac{x}{2q_1}) + \mathbb{P}(\|\mathcal{E}_n^{k,\text{trunc}} - \mathcal{F}_n^{k,\text{trunc}}\| \geq (\frac{x}{2q_1})^{1/q_2}) \\
\leq \quad & \exp\{-\frac{n^2(x/2q_1)^2}{n\psi_2(n) + \psi_3(n)n(x/2q_1)/3}\} + \exp\{-\frac{n^2(x/2q_1)^{2/q_2}}{n\psi_2(n) + \psi_3(n)n(x/2q_1)^{1/q_2}/3}\} \\
\leq \quad & 2\exp\{-n\phi_n(x)\}, \qquad (32)
\end{aligned}$$

where we define

$$\phi_n(y) \quad = \quad \min\left\{\frac{(y/2q_1)^2}{\psi_2(n) + \psi_3(n)(y/2q_1)/3}, \frac{(y/2q_1)^{2/q_2}}{\psi_2(n) + \psi_3(n)(y/2q_1)^{1/q_2}/3}\right\}. \qquad (33)$$

10 Proof of Main Results

To start with, we first prove Proposition 2.

Proof of Proposition 2. We first define event $H_1 = \{\delta_1 < \nu_1 < \delta_1 + L\}$ with $L = w\lceil \frac{K}{v} \rceil$. We then have

$$\begin{aligned} & \mathbb{E}\Big[\nu_1 - \delta_1\Big] \\ = \; & \mathbb{E}\Big[(\nu_1 - \delta_1)\mathbb{I}_{H_1}\Big] + \mathbb{E}\Big[(\nu_1 - \delta_1)\mathbb{I}_{H_1^c}\Big] \\ = \; & \mathbb{E}\Big[(\nu_1 - \delta_1)|H_1\Big]\mathbb{P}(H_1) + \mathbb{E}\Big[(\nu_1 - \delta_1)\mathbb{I}_{H_1^c}\Big] \\ \leq \; & \mathbb{E}\Big[(\nu_1 - \delta_1)|H_1\Big]\mathbb{P}(\nu_1 \leq \delta_1 + L) + T\mathbb{P}(\nu_1 > \delta_1 + L) \\ \leq \; & L(1 - \text{prob}_1(\beta)) + T\text{prob}_1(\beta). \end{aligned} \tag{34}$$

This completes the proof of Proposition 2. □

10.1 Pseudo Regret Decomposition

We next prove the pseudo regret decomposition.

Proof of Proposition 3. By the stability property of the risk function, we have for any $F_1, F_2 \in \mathcal{D}^\Lambda$,

$$\begin{aligned} |U(F_1) - U(F_2)| &\leq q_1\Big(\|F_1 - F_2\| + \|F_1 - F_2\|^{q_2}\Big) \\ &= q_1\Big(1 + \|F_1 - F_2\|^{q_2 - 1}\Big)\|F_1 - F_2\| \\ &\leq q_1\Big(1 + \mathbb{D}^{q_2 - 1}\Big)\|F_1 - F_2\| \\ &= G\|F_1 - F_2\|. \end{aligned}$$

Using above result, we have

$$\begin{aligned} \mathcal{R}_{\boldsymbol{\pi}}(T) &= \mathbb{E}\Big[\Big|U(\mathcal{F}_T^{\boldsymbol{\pi}^*}) - U(\mathcal{F}_T^{\boldsymbol{\pi}})\Big|\Big] \\ &\leq G\mathbb{E}\Big[\|\mathcal{F}_T^{\boldsymbol{\pi}^*} - \mathcal{F}_T^{\boldsymbol{\pi}}\|\Big] \\ &\leq \frac{G}{T}\mathbb{E}\Big[\sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{M}}\sum_{t=\delta_{i-1}+1}^{\delta_i} \|F^i_{\pi_t^*(i)} - F^i_{\pi_t}\|\mathbb{1}\Big\{\pi_t = k;\; k \notin \pi_t^*(i)\Big\}\Big] \\ &\leq \frac{G}{T}\mathbb{E}\Big[\sum_{k\in\mathcal{K}}\Delta_k\rho\sum_{i\in\mathcal{M}}\sum_{t=\delta_{i-1}+1}^{\delta_i} \mathbb{1}\Big\{\pi_t = k;\; k \notin \pi_t^*(i)\Big\}\Big] \\ &= \frac{\rho G}{T}\sum_{k\in\mathcal{K}}\Delta_k\mathbb{E}[\eta_k(T)] \end{aligned}$$

□

10.2 Performance of RS-SCB

Let $u_{i,k}$ denote the minimal positive solution to the following equation,

$$(U(F^i_{\pi^*(i)}) - U(F^i_k))/2 = \phi_u^{-1}(\frac{\alpha\log T}{u}) + \bar{\psi}_1(u). \tag{35}$$

Here $u_{i,k}$ can be interpreted as the length of phase required to figure out the best arm in each regime i.

To prove Theorem 1, we first prove the situation when there is no change point.

Proposition 4. The pseudo regret of Algorithm 2 when there are no change points $M = 1$, is upper bounded as

$$\mathcal{R}_{\boldsymbol{\pi}}(T) \quad \leq \quad \frac{\rho G}{T}\left(\sum_{k\in\mathcal{K}} u_{min,k}\Delta_k + \Delta_k\right) + \frac{\rho G \zeta}{K}\sum_{k\in\mathcal{K}}\Delta_k,$$

where $u_{min,k}$ is defined in Eq. (36).

Proof. We have the following

$$\mathcal{R}_{\boldsymbol{\pi}}(T) = \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu \leq T\Big\}\Big] + \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu > T\Big\}\Big]$$

From propositions, choosing $\beta = \chi/3$ satisfying equation (21), we have that the probability of false alarm $P(\nu \leq T) \leq \frac{1}{T}$. Using a trivial bound $\mathcal{R}_{\boldsymbol{\pi}}(T) \leq \rho G \sum_{k\in\mathcal{K}}\Delta_k$, we have

$$\mathcal{R}_{\boldsymbol{\pi}}(T) \leq \frac{\rho G}{T}\sum_{k\in\mathcal{K}}\Delta_k + \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu > T\Big\}\Big].$$

Note that $\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu > T\Big\}\Big]$ is the expected pseudo regret when there is no false alarm. Using the result in *truncation exploration and exploitation* (see Eq. (40)), we have

$$\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu > T\Big\}\Big] \quad \leq \quad \frac{\rho G}{T}\left(u_{min,k}\Delta_k\right) + \frac{\rho G \zeta}{K}\sum_{k\in\mathcal{K}}\Delta_k.$$

The result follows. □

Proof of Theorem 1*:* Let $H_i = \{\delta_i < \nu_i \leq \delta_i + L\}$, where $i = \{1, 2, \cdots M\}$ and $L = w\lceil\frac{K}{\zeta}\rceil$. We have the following decomposition for the pseudo regret by viewing the restart after every detection as a renewal process,

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\pi}}(T) &= \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu_1 \leq \delta_1\Big\}\Big] + \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu_1 > \delta_1\Big\}\Big] \\
&\leq \frac{\rho G}{T}\sum_{k\in\mathcal{K}}\Delta_k + \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T) - \mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big] \\
&+ \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\mathbb{1}\Big\{\nu_1 > \delta_1\Big\}\Big] \\
&\leq \underbrace{\frac{\rho G}{T}\left(u_{min,k}\Delta_k + \Delta_k\right)}_{\mathcal{V}} \\
&+ \frac{\rho G \zeta \delta_1}{TK}\sum_{k\in\mathcal{K}}\Delta_k + \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T) - \mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big].
\end{aligned}$$

Consider the decomposition of $\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T) - \mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big]$. We have

$$\begin{aligned}
\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T) - \mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big] \leq \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T) - \mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big|H_1\Big] \\
+ \rho G \sum_{k\in\mathcal{K}}\Delta_k \cdot (1 - P(H_1)).
\end{aligned}$$

We have for the restarting RS-UCB algorithm

$$\begin{aligned}
\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)-\mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big] &\le \mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)-\mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big|H_1\Big]+\frac{\rho G}{T}\sum_{k\in\mathcal{K}}\Delta_k\\
&\le \frac{\rho G}{T}\sum_{k\in\mathcal{K}}\Delta_k+\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)-\mathcal{R}_{\boldsymbol{\pi}}(\nu_1)\Big|H_1\Big]+\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(\nu_1)-\mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big|H_1\Big]\\
&\le \frac{\rho G}{T}\sum_{k\in\mathcal{K}}\Delta_k+\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T-\delta_1)\Big]+\frac{\rho G}{T}\mathbb{E}\Big[\nu_1-\delta_1\Big|H_1\Big]\sum_{k\in\mathcal{K}}\Delta_k.
\end{aligned}$$

Using the bound of detection delay, we have

$$\begin{aligned}
\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T)-\mathcal{R}_{\boldsymbol{\pi}}(\delta_1)\Big] \le \frac{\rho G}{T}\sum_{k\in\mathcal{K}}\Delta_k+\frac{\rho G}{T}\max_{k\in\mathcal{K}}\Delta_k\cdot L\\
+\,\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T-\delta_1)\Big],
\end{aligned}$$

where we use the trivial fact that $\mathbb{E}\Big[\nu_1-\delta_1\Big|H_1\Big]\le L$.

Now, we can perform the same decomposition for $\mathbb{E}\Big[\mathcal{R}_{\boldsymbol{\pi}}(T-\delta_1)\Big]$, and this can be done at most M times. Therefore, we obtain

$$\mathcal{R}_{\boldsymbol{\pi}}(T)\le\frac{\rho G\zeta}{K}\sum_{k\in\mathcal{K}}\Delta_k+\mathcal{V}M+\frac{\rho GM}{T}\sum_{k\in\mathcal{K}}\Delta_k+M\frac{2w\rho GK}{\zeta T}\max_{k\in\mathcal{K}}\Delta_k.$$

Now we define the events that

$$\begin{aligned}
V_k^n&=\Big\{U(\mathcal{E}^{k,\mathrm{trunc}}_{n_k(t)})>U(\mathcal{F}^{k}_{n_k(t)})+\phi^{-1}_{n_k(t)}(\frac{\alpha\log t}{n_k(t)})+\bar{\psi}_1(n_k(t))\Big\},\\
V_{k^*}^t&=\Big\{U(\mathcal{E}^{k^*,\mathrm{trunc}}_{n_k^*(t)})<U(\mathcal{F}^{k^*}_{n_{k^*}(t)})-\phi^{-1}_{n_k(t)}(\frac{\alpha\log t}{n_k^*(t)})-\bar{\psi}_1(n_k^*(t))\Big\},
\end{aligned}$$

where we recall $\bar{\psi}_1(n)=q_1(\psi_1(n)+\psi_1(n)^{q_2})$.

It is not hard to see that

$$V_k^t\subset\Big\{U(\mathcal{E}^{k,\mathrm{trunc}}_{n_k(t)})>U(\mathcal{F}^{k,\mathrm{trunc}}_{n_k(t)})+\phi^{-1}_{n_k(t)}(\frac{\alpha\log t}{n_k(t)})\Big\}$$

and

$$V_{k^*}^t\subset\Big\{U(\mathcal{E}^{k^*,\mathrm{trunc}}_{n_{k^*}(t)})<U(\mathcal{F}^{k^*,\mathrm{trunc}}_{n_{k^*}(t)})-\phi^{-1}_{n_k(t)}(\frac{\alpha\log t}{n_k^*(t)})\Big\}.$$

Then, we could compute that

$$\begin{aligned}
&\mathbb{P}(V_k^t)\\
\le\;&\mathbb{P}(\Big\{U(\mathcal{E}^{k,\mathrm{trunc}}_{n_k(t)})>U(F^{k,\mathrm{trunc}}_{n_k(t)})+\phi^{-1}_{n_k(t)}(\frac{\alpha\log t}{n_k(t)})\Big\})\\
\le\;&\mathbb{P}(\bigcup_{1\le s\le t}\Big\{U(\mathcal{E}^{k,\mathrm{trunc}}_{s})>U(\mathcal{F}^{k,\mathrm{trunc}}_{s})+\phi_s^{-1}(\frac{\alpha\log t}{s})\Big\})\\
\le\;&\sum_{s=1}^{t}2\exp\{-s\phi_s(\phi_s^{-1}(\frac{\alpha\log t}{s}))\}\\
\le\;&\sum_{s=1}^{t}\frac{2}{t^{\alpha}}\le\frac{2}{t^{\alpha-1}}.
\end{aligned}$$

Choose an integer u large enough such that

$$\Delta_k/2\ge\phi_u^{-1}(\frac{\alpha\log T}{u})+\bar{\psi}_1(u). \tag{36}$$

We denote the such $u_{min,k}$ be the minimal positive solution to the inequality Eq. (36). ($u_{min,k}$ becomes $u_{i,k}$ when Δ_k is replaced by $U(F^i_{\boldsymbol{\pi}^*(i)}) - U(F^i_k)$.)

For example, when we take $\psi_1(t) = (t/\alpha \log \frac{1}{\delta})^{-\epsilon/(1+\epsilon)}$, $\psi_2(t) = (t/\alpha \log \frac{1}{\delta})^{(1-\epsilon)/(1+\epsilon)}$ and $\psi_3(t) = (t/\alpha \log \frac{1}{\delta})^{1/(1+\epsilon)}$, it leads to

$$\begin{aligned} u_{min,k} \quad \leq \quad & \max\{ \frac{\alpha \log T}{(\Delta_k/12q_1)^{1+\epsilon}/\epsilon}, \\ & \frac{\alpha \log T}{(\Delta_k/12q_1)^{1+\epsilon}/q_2\epsilon}, \\ & (\frac{6q_1}{\Delta_k})^{\frac{1+\epsilon}{\epsilon}}, (\frac{6q_1}{\Delta_k})^{\frac{1+\epsilon}{q_2\epsilon}} \}. \end{aligned} \tag{37}$$

By such choice of $u_{min,k}$, we can see that

$$\{\pi_t = k\} \cap \{n_k(t) \geq u_{min,k}\} \subset V_k^t \cup V_{k^*}^t.$$

If not, $\{\pi_t = k\} \cap \{n_k(t) \geq u_{min,k}\} \cap (V_k^t \cup V_{k^*}^t)^c \neq \emptyset$, we then know

$$\begin{aligned} & U(\mathcal{E}^{k^*,\text{trunc}}_{n_{k^*}(t)}) + \phi^{-1}_{n_k(t)}(\frac{\alpha \log t}{n_{k^*}(t)}) + \bar{\psi}_1(n_{k^*}(t)) \\ > \quad & U(\mathcal{F}^{k^*}_{n_{k^*}(t)}) \\ = \quad & U(\mathcal{F}^{k}_{n_k(t)}) + \Delta_k \\ \geq \quad & U(\mathcal{F}^{k}_{n_k(t)}) + 2(\phi^{-1}_{n_k(t)}(\frac{\alpha \log t}{n_k(t)}) + \bar{\psi}_1(n_k(t)) \\ \geq \quad & U(\mathcal{E}^{k,\text{trunc}}_{n_k(t)}) + \phi^{-1}_{n_k(t)}(\frac{\alpha \log t}{n_k(t)}) + \bar{\psi}_1(n_k(t). \end{aligned} \tag{38}$$

This leads to the contradiction that $\pi_t \neq k$.

Finally, we have that

$$\begin{aligned} \mathbb{E}[n_k(T)] \quad = \quad & \mathbb{E}[\sum_{t=1}^{T} \mathbf{1}\{\pi_t = k\}] \\ \leq \quad & u_{min,k} + \mathbb{E}[\sum_{u+1}^{T} \mathbf{1}\{\pi_t = k \cap n_k(t) \geq u_{min,k}\}] \\ \leq \quad & u_{min,k} + \sum_{t=K+1}^{T} \frac{2}{t^{\alpha-1}} \\ \leq \quad & u_{min,k} + \frac{2}{(\alpha-1)K^{\alpha-2}} \\ \leq \quad & u_{min,k} + 2. \end{aligned} \tag{39}$$

When there is no change point, the pseudo regret is upper bounded by

$$\mathbb{E}[\mathcal{R}_{\boldsymbol{\pi}}(T)\mathbb{1}\Big\{\nu > T\Big\}] \leq \frac{\rho G}{T} \sum_{k \in \mathcal{K}} u_{min,k}\Delta_k + \frac{\rho G v}{K} \sum_{k \in \mathcal{K}} \Delta_k. \tag{40}$$

This completes the proof.

Proof of Corollary 2. For the simplicity of proof, we can assume that $\Delta_{i,k} := U(F^i_{\pi^*(i)}) - U(F^i_k)$ has the same order for $i \in \{1, \ldots, M\}$. Based on Eq. 16, it is straightforward to find that

$$u_k \leq c(\frac{\log T}{\Delta_k})^{\frac{1+\epsilon}{\epsilon}},$$

when we choose $\psi_1(n) = n^{-\epsilon/(1+\epsilon)}$, $\psi_2(n) = n^{\frac{1-\epsilon}{1+\epsilon}}$ and $\psi_3(n) = n^{\frac{1}{1+\epsilon}}$. Here c is a universal constant.

Let N_k be the pulling number of arm k. Then together with Eq. 17, we know that

$$\mathcal{R}_{\mathrm{RS-SCB}}(T) \leq \min\{M\Big(\mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3\Big) + \frac{\rho G \zeta}{K} \sum_{k \in \mathcal{K}} \Delta_k, \frac{\rho G}{T} \sum_k N_k \Delta_k\}. \tag{41}$$

By expanding $\mathcal{C}_1$ - $\mathcal{C}_3$ and plugging in $\zeta = \sqrt{MwK/T}$, we have

$$\mathcal{R}_{\mathrm{RS-SCB}}(T) \quad \leq \quad \frac{\rho G}{T} \sum_k \min\{M u_k, N_k\} \Delta_k + O(\rho G \sqrt{\frac{MwK}{T}}). \tag{42}$$

The right hand side of Eq. (42) can be optimized with respect to Δ_k and it gives

$$\begin{aligned} \mathcal{R}_{\mathrm{RS-SCB}}(T) \quad &\leq \quad c' \frac{\rho G}{T} \sum_k \{(\log T) M^{\frac{\epsilon}{1+\epsilon}} N_k^{\frac{1}{1+\epsilon}}\} + O(\rho G \sqrt{\frac{MwK}{T}}) \\ &\leq \quad c' \frac{\rho G}{T} (\log T)(MK)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} + O(\rho G \sqrt{\frac{MwK}{T}}), \end{aligned} \tag{43}$$

where the last inequality in Eq. (43) uses the fact that $x^{1/(1+\epsilon)}$ is concave and Jenson's inequality. Hence we conclude that

$$\mathcal{R}_{\mathrm{RS-SCB}}(T) = O(\rho G (\log T) (\frac{MK}{T})^{\epsilon/(1+\epsilon)}).$$

□

Proof of Theorem 2 is quite similar as that of Theorem 1. Hence we omit here.

11 Comments on Lower Bound

In this section, we provide a discussion on the lower bound. We assume the following structural assumptions on risk function.

Assumption 6 (U-structure). *Suppose the risk function U is a composite function of several linear measures, i.e, satisfies $U := H(U_1, \dots, U_L) = \sum_{l=1}^{L_0} g_l(U_l) - \sum_{l=L_0+1}^{L} h_l(U_l)$, where g_l's are convex functions and h_l's are concave functions.*

Theorem 3 (Instance-dependent Lower Bound on Pseudo Regret). *Suppose Assumptions 1-3 and 6 hold. For any admissible policy π, we can always find a K-armed problem with sub-optimality gap Δ and at most $M-1$ change points such that*

$$\mathbb{E}[U(\mathcal{F}_T^{\boldsymbol{\pi}^*}) - U(\mathcal{F}_T^{\boldsymbol{\pi}})] \geq \Omega(\Delta \frac{\sqrt{MK}}{\sqrt{T}}). \tag{44}$$

The above theorem says that, for any admissible policy $\boldsymbol{\pi}$, its worst instance-dependent bound, under suitable structural assumption on the risk, is at least $\sqrt{MK/T}$ if we treat sub-optimality gap is fixed and free of T.

Proof of Theorem 3. First, for the simplicity, we assume the best arm does not change over the whole time horizon. Let k^* be the arm maximizing $U(F_k)$ and N_k be the number of times that arm k has been pulled. We then know the oracle policy maximizing $U(\mathcal{F}_T^{\boldsymbol{\pi}})$ is the one always choosing arm k^*.

For any $\boldsymbol{\pi}$, we can have the following inequality.

$$\begin{aligned}
& U(\mathcal{F}_T^{\boldsymbol{\pi}^*}) - U(\mathcal{F}_T^{\boldsymbol{\pi}}) \\
= \; & H(U_1(\mathcal{F}_T^{\boldsymbol{\pi}^*}), \dots, U_L(\mathcal{F}_T^{\boldsymbol{\pi}^*})) - H(U_1(\mathcal{F}_T^{\boldsymbol{\pi}}), \dots, U_L(\mathcal{F}_T^{\boldsymbol{\pi}})) \\
= \; & H(U_1(F_{k^*}), \dots, U_L(F_{k^*})) - H(U_1(\mathcal{F}_T^{\boldsymbol{\pi}}), \dots, U_L(\mathcal{F}_T^{\boldsymbol{\pi}})) \\
= \; & \sum_{l=1}^{L_0} g_l(U_l(F_{k^*})) - \sum_{l=L_0+1}^{L} h_l(U_l(F_{k^*})) - \Big(\sum_{l=1}^{L_0} g_l(U_l(\mathcal{F}_T^{\boldsymbol{\pi}})) - \sum_{l=L_0+1}^{L} h_l(U_l(\mathcal{F}_T^{\boldsymbol{\pi}}))\Big) \\
= \; & \sum_{l=1}^{L_0} g_l(U_l(F_{k^*})) - \sum_{l=L_0+1}^{L} h_l(U_l(F_{k^*})) \\
& -\Big(\sum_{l=1}^{L_0} g_l(U_l(\sum_{i=1}^{K} \frac{N_k}{T} F_k)) - \sum_{l=L_0+1}^{L} h_l(U_l(\sum_{k=1}^{K} \frac{N_k}{T} F_k))\Big) \\
\geq \; & \sum_{l=1}^{L_0} g_l(U_l(F_{k^*})) - \sum_{l=L_0+1}^{L} h_l(U_l(F_{k^*})) \\
& -\Big(\sum_{l=1}^{L_0} \sum_{i=1}^{K} \frac{N_k}{T} g_l(U_l(F_k)) - \sum_{l=L_0+1}^{L} \sum_{k=1}^{K} \frac{N_k}{T} h_l(U_l(F_k))\Big) \qquad (45) \\
= \; & \sum_{l=1}^{L_0} \sum_{k=1}^{K} \frac{N_k}{T} [g_l(U_l(F_{k^*})) - g_l(U_l(F_k))] - \sum_{l=L_0+1}^{L} \sum_{k=1}^{K} \frac{N_k}{T} [h_l(U_l(F_{k^*})) - h_l(U_l(F_k))] \\
= \; & \sum_{k=1}^{K} \frac{N_k}{T} \Big(\sum_{l=1}^{L_0} g_l(U_l(F_{k^*})) - g_l(U_l(F_k)) - [\sum_{l=L_0+1}^{L} h_l(U_l(F_{k^*})) - h_l(U_l(F_k))]\Big) \\
= \; & \sum_{k=1}^{K} \frac{N_k}{T} \Delta_k. \qquad (46)
\end{aligned}$$

Here, Eq. (45) holds due to the convexity of g_l's, concavity of h_l's and linearity of U_l's. Δ_k is defined as $U(F_{k^*}) - U(F_k)$. By using the treatment below on the lower bound of $\mathbb{E}[N_k]$ for $k \neq k^*$,

$$\mathbb{E}[U(\mathbb{E}[\mathcal{E}_T^{\boldsymbol{\pi}^*}]) - U(\mathbb{E}[\mathcal{E}_T^{\boldsymbol{\pi}}])] \geq \sum_{k=1}^{K} \frac{\mathbb{E}[N_k]}{T} \Delta_k.$$

Dealing with $\mathbb{E}[N_k]$:

We consider a K-armed bandit setting with M stationary regimes (i.e. at most $M-1$ change points). We split the entire time interval $[0, T]$ into $\lfloor\sqrt{\frac{T}{KM}}\rfloor$ non-overlapping intervals I_l's. Each interval I_l has length $\sqrt{KTM}$. Furthermore, we split each I_l into M sub intervals with each sub intervals having length $\sqrt{\frac{KT}{M}}$. We let $N_{l,m}(k)$ be the number of times that arm k is played during m-th sub-interval of I_l. Hence $N_k = \sum_{l,m} N_{l,m}(k)$.

If there is an arm $k \in [K]$ such that there exists a set $\mathcal{A}_k$ containing at least $\frac{1}{2}M$ pairs of (l, m) such that $\mathbb{E}_{\pi,1}[N_{l,m}(k) = o(1)]$, where π is an arbitrary policy and subscript "1" indicates the setting that arm 1 is the optimal arm over the entire time horizon. Then we can construct another setting "2" such that arm k is the optimal arm during those $\frac{1}{2}M$ sub-intervals in $\mathcal{A}_k$. and arm 1 is the optimal arm in the remaining intervals. It is easy to check that this setting contains $M-1$ change points. In addition, we assume that the reward distribution of arm k changes from F_k to F'_k such that $U(F'_k) - U(F_1) =: \Delta > 0$ and $\alpha := KL(F_2 \| F'_2) > 0$.

By Lemma A.1 in Auer et.al. (2002b), we have that

$$\mathbb{E}_{\pi,2}[N_{l,m}(k)] \le E_{\pi,1}[N_{l,m}(k)] + \frac{\tau_0}{2}\sqrt{\alpha \mathbb{E}_{\pi,1}[N_{l,m}(k)]}, \tag{47}$$

where $\tau_0 = \sqrt{\frac{KT}{M}}$. In other words, $\mathbb{E}_{\pi,2} = o(\tau_0)$. Therefore, the regret of setting 2 is at least

$$\sum_{(l,m)\in\mathcal{A}_k} \mathbb{E}[\tau_0 - N_{l,m}(k)] \ge \frac{1}{2}\frac{M}{2}\tau_0 = \frac{1}{4}\sqrt{MKT}.$$

On the other hand, if for all $k \in [K]$, there exist at most $\frac{1}{2}M$ pairs of (l, m) such that $\mathbb{E}_{\pi,1}[N_{l,m}(k)] = o(1)$. We let $\mathcal{B}_k$ be the set containing all sub-intervals with $\mathbb{E}_{\pi,1}[N_{l,m}(k)] \ge c$. ($c$ is some small positive constant). Thus $|\mathcal{B}_k| \ge \sqrt{\frac{T}{KM}}M - \frac{1}{2}M \ge \sqrt{\frac{T}{KM}}M/2$. Then we can compute the lower bound of setting 1 as follows.

$$\sum_{(l,m)\in\mathcal{B}_k}\sum_{k\in[K]} \mathbb{E}_{\pi,1}[N_{l,m}(k)] \ge K\frac{M}{2}\sqrt{T/KM}c = \frac{c}{2}\sqrt{MKT}.$$

In a summary, the unnormalized instance-dependent bound is at least $\Omega(\sqrt{MKT}\Delta)$. This completes the proof. □

Theorem 4 (Minimax Lower Bound on Pseudo Regret). *Suppose Assumptions 1, 2, 4 and 6 hold. For any admissible policy π, we can always find a K-armed problem with at most $M-1$ change points such that*

$$\mathbb{E}[U(\mathcal{F}_T^{\boldsymbol{\pi}^*}) - U(\mathcal{F}_T^{\boldsymbol{\pi}})] \ge \Omega((MK/T)^{\frac{\epsilon}{1+\epsilon}}). \tag{48}$$

Proof of Theorem 4. In this proof, we consider to construct a set of non-stationary Bernoulli bandits. Define $\nu_1 = (1-\gamma^{1+\epsilon})\delta_0 + \gamma^{1+\epsilon}\delta_{1/\gamma}, \nu_2 = (1 + \Delta\gamma - \gamma^{1+\epsilon})\delta_0 + (\gamma^{1+\epsilon} - \Delta\gamma)\delta_{1/\gamma}$. Set $\mathcal{M}_0 = \{\{F_1, \ldots, F_K\}$: there is one $F_k = \nu_1$, rest are equal to $\nu_2\}$ over T/M rounds. (Here we assume T/M is an integer without loss of generality.) We then set $\mathcal{M}_{non-stationary} = \underbrace{\mathcal{M}_0 \oplus \ldots \oplus M_0}_{m \text{ times}}$, which is a concatenation of M bandits with $\frac{T}{M}$ rounds and each one belongs to $\mathcal{M}_0$.

By choosing $\Delta = (KM/T)^{\epsilon/(1+\epsilon)}$ and $\gamma = (2\Delta)^{1/\epsilon}$, we have that the regret over each bandit in $\mathcal{M}_0$ is at least $0.01K^{\frac{\epsilon}{1+\epsilon}}(T/M)^{\frac{1}{1+\epsilon}}$ by Theorem 2 in Bubeck et al. (2013). Finally, the total un-normalized regret is at least $0.01K^{\frac{\epsilon}{1+\epsilon}}(T/M)^{\frac{1}{1+\epsilon}}M = \Omega((KM)^{\frac{\epsilon}{1+\epsilon}}(T)^{\frac{1}{1+\epsilon}})$. This completes the proof. □