Catoni-Style Change Point Detection for Regret Minimization in Piecewise-Stationary Heavy-Tailed Bandits

Gianmarco Genalti
Politecnico di Milano

Sujay Bhatt
J.P. Morgan AI Research

Nicola Gatti
Politecnico di Milano

Alberto Maria Metelli
Politecnico di Milano

Abstract

Regret minimization in stochastic non-stationary bandits gained popularity over the last decade, as it can model a broad class of real-world problems, from advertising to recommendation systems. Existing literature relies on various assumptions about the reward-generating process, such as Bernoulli or subgaussian rewards. However, in settings such as finance and telecommunications, *heavy-tailed* distributions naturally arise. In this work, we tackle the heavy-tailed piecewise-stationary bandit problem. Heavy-tailed bandits, introduced by Bubeck et al., 2013, operate on the minimal assumption that the finite absolute centered moments of maximum order $1+\epsilon$ are uniformly bounded by a constant $v < +\infty$, for some $\epsilon \in (0, 1]$. We focus on the most popular non-stationary bandit setting, *i.e.*, the piecewise-stationary setting, in which the mean of reward-generating distributions may change at unknown time steps. We provide a novel Catoni-style change-point detection strategy tailored for heavy-tailed distributions that relies on recent advancements in the theory of sequential estimation, which is of independent interest. We introduce `Robust-CPD-UCB`, which combines this change-point detection strategy with optimistic algorithms for bandits, providing its regret upper bound and an impossibility result on the minimum attainable regret for any policy. Finally, we validate our approach through numerical experiments on synthetic and real-world datasets.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

In a Multi-Armed Bandit (MAB, for short, Lattimore and Szepesvári, 2020), a decision-maker (also called *learner*) is faced with a sequence of repeated decisions among a fixed number of options (or *actions* or *arms*), observing a reward drawn from a probability distribution after each decision. MABs gained popularity over the last two decades, as they allow for strong theoretical guarantees over algorithms' performance while maintaining the model's generality. However, the most traditional MAB model relies on demanding assumptions that are rarely met in the real world. The majority of the research effort in the MAB literature focused on progressively overcoming these limitations to cover richer scenarios.

In this work, we focus on a broad class of problems that relaxes, at the same time, two core assumptions of the standard MAB problem: *heavy-tailed non-stationary* MABs. Our framework allows for a general class of reward-generating probability distributions without relying on parametric assumptions and with a possibly infinite variance, called heavy-tailed distributions. This setting gained popularity over the last decade due to its applications in finance and telecommunications, and it overcomes the assumption of subgaussian reward distributions, which is customary in the MAB literature. In such application domains, the assumption that reward-generating distributions are fixed along the whole time horizon is too limiting. It is natural to consider settings, such as finance, characterized by non-stationary reward processes. We address, with a single algorithm, named `Robust-CPD-UCB`, the problem of learning in non-stationary environments where the noise of the observations can be heavy-tailed. We prove theoretical guarantees on the performance of `R-CPD-UCB` and show that they are nearly optimal under some mild assumptions. To the best of the authors' knowledge, this is the first work to address the problem of regret minimization in non-stationary bandits under infinite-variance reward distributions. In particular, we face the technical challenge of developing

the first change-point detection strategy with proven theoretical guarantees for such types of distributions.

2 PROBLEM FORMULATION

In this section, we recall the definitions of heavy-tailed and piecewise-stationary bandit. Then, we introduce the heavy-tailed piecewise-stationary bandits, the focus of this work. We formally define the problem of regret minimization and provide a novel regret lower bound.

2.1 Bandit Settings

Every round $t \in [T] \coloneqq \{1, \dots, T\}$, a decision $I_t \in [K]$ is undertaken (possibly at random) and a reward $X_{I_t,t}$ is sampled from a probability distribution ν_{I_t}. We call the set $\boldsymbol{\nu} = \{\nu_i\}_{i \in [K]}$ of reward-generating distributions an *instance* of the MAB (from now on, just MAB). Most of the literature deals with reward-generating distributions either *subgaussian* or with bounded supports.

Heavy-Tailed Bandits. In *heavy-tailed* bandits (Bubeck et al., 2013), the probability distributions $\{\nu_i\}_{i \in [K]}$ are *heavy-tailed*. In this work, we use the same definition of heavy-tailed MAB (HT MAB, for short).

Definition 2.1 (Heavy-Tailed MAB)**.** Let $X \sim \nu$ be a random variable with support on $\mathbb{R}$. Then, we call X a *heavy-tailed random variable* if it satisfies

$$\mathbb{E}_\nu[|X - \mathbb{E}_\nu[X]|^{1+\epsilon}] \leq v, \tag{1}$$

for $\epsilon \in (0, 1]$ and $v \in \mathbb{R}^+$. Let $\boldsymbol{\nu}$ be a MAB. Then, if $\boldsymbol{\nu} \in \mathcal{H}^K_{(v,\epsilon)}$, where $\mathcal{H}_{(v,\epsilon)}$ is the set of probability distributions satisfying Equation (1), we call $\boldsymbol{\nu}$ a *heavy-tailed bandit* (HT MAB, for short).

Note that Equation (1) implies that the variance of the rewards-generating distributions may be infinite (when $\epsilon < 1$). Most of the technical tools employed for subgaussian rewards are ineffective for HT MABs. We address readers to Genalti et al. (2024a) for a recent literature review on HT MABs.

Piecewise-Stationary Bandits. In standard MABs, the reward-generating distributions are assumed to never change during learning. In *non-stationary bandits*, instead, the reward-generating distributions are dynamic in time, *i.e.*, the rewards of the same arm are sampled from different distributions depending on the pull time $t \in [T]$. However, if there is no constraint on how many times the distributions may change, then the problem may quickly become non-tractable. Thus, in this work, we consider the most popular non-stationary MAB setting, *i.e.*, the *piecewise-stationary* bandit (PS MAB, for short) from Yu and Mannor (2009), where the distributions of rewards remain constant for a certain period, called *epoch*, and then abruptly change at some unknown time points, called *breakpoints*. We assume that the total number of breakpoints $\Upsilon \in [T]$ is fixed before the trial. We define a PS MAB as follows.

Definition 2.2 (Piecewise-Stationary Bandit)**.** Let $\{\boldsymbol{\nu}^{(j)}\}_{j \in [\Upsilon]}$ be a set of Υ MABs. Then, let $\{t_c^{(j)}\}_{j \in [\Upsilon]} \subset [T]$ be the set of breakpoints and call E_j the set of indices $\{t_c^{(j-1)}, \dots, t_c^{(j)}\}$, where $t_c^{(0)} = 0$ and $t_c^{(\Upsilon+1)} = T$, by convention. If $\nu_i^{(j)}$ is the reward-generating distribution of arm $i \in [K]$ when $t \in E_j$, then $(\{\boldsymbol{\nu}^{(j)}\}_{j \in [\Upsilon]}, \{t_c^{(j)}\}_{j \in [\Upsilon]})$ defines a *piecewise-stationary bandit* (PS MAB, for short).

E_j is called the j-th epoch of the PS MAB, and $t_c^{(j)}$ to the j-th breakpoint. To simplify notation, we define $\mu_i^{(j)} \coloneqq \mathbb{E}_{\nu_i^{(j)}}[X_{i,t}]$ as the mean of the reward-generating distribution of action i during epoch j. Note that the reward-generating distribution of an action is fixed during an epoch, and so is the mean reward (and every other distribution parameter). We call $\delta_i^{(j)} \coloneqq |\mu_i^{(j)} - \mu_i^{(j-1)}|$ the magnitude of the change in the mean of arm $i \in [K]$ from epoch E_{j-1} to the next one, E_j. By convention, $E_0 = \emptyset$ and $\delta_i^{(0)} = \infty$ for every $i \in [K]$. In Appendix B, we review the literature on PS MABs.

Piecewise Non-stationary Heavy-Tailed Bandits. The general definition of piecewise non-stationary MABs allows for any family of reward-generating distributions, including heavy-tailed ones. In this work, we deal with piecewise non-stationary bandits where the reward-generating distributions satisfy Equation (1). We call this setting the *heavy-tailed piecewise-stationary* setting (HTPS MAB, for short).

Definition 2.3 (Heavy-Tailed Piecewise-Stationary Bandits)**.** Let $(\{\boldsymbol{\nu}^{(j)}\}_{j \in [\Upsilon]}, \{t_c^{(j)}\}_{j \in [\Upsilon]})$ be a PS bandit. If $\boldsymbol{\nu}^{(j)} \in \mathcal{H}^K_{(v,\epsilon)}$ for every $j \in [\Upsilon]$, we call it a heavy-tailed piecewise-stationary bandit (HTPS MAB, for short). We denote the set of such HTPS MABs as $\mathcal{B}_{(v,\epsilon,\Upsilon)}$.

Definition 2.3 introduces the novel bandit setting, as the intersection between HT and PS MABs. To the best of the authors' knowledge, this setting has not been studied in previous literature.

2.2 Learning Goal

A policy is a (possibly randomized) map $\pi(t) : \mathcal{F}_{t-1} \mapsto I_t$ that receives the filtration up to time $t-1$ (composed of past actions and rewards) and returns an

action $I_t \in [K]$ to play. We define $\Delta_i^{(j)}$ as the *sub-optimality gap* of arm $i \in [K]$ during epoch $j \in [\Upsilon]$, *i.e.*, $\Delta_i^{(j)} := |\max_{k\in[K]} \mu_k^{(j)} - \mu_i^{(j)}|$ and $N_{i,j}^{\pi}(t)$ as the number of times action i has been chosen during epoch j by policy π up to time $t \in [T]$. The goal of a learner is to minimize the *expected cumulative regret* $\mathbb{E}[R^{\pi}(T)]$, *i.e.*, the cumulative performance gap w.r.t. to the best *policy* over a learning horizon.

Definition 2.4 (Expected Cumulative Regret)**.** Given a policy π, we define the expected cumulative regret of π as:

$$\mathbb{E}[R^{\pi}(T)] = \sum_{j\in[\Upsilon]} \sum_{i\in[K]} \Delta_i^{(j)} \mathbb{E}[N_{i,j}^{\pi}(T)],$$

where the expectation accounts for both the randomness of reward generation and policy π.

This performance index is also called *dynamic regret* and is the standard choice for PS MABs. The optimal policy corresponds to choosing the best action in every epoch $j \in [\Upsilon]$, *i.e.*, $i_j^* \in \arg\max_{i\in[K]} \mu_k^{(j)}$. We also define some quantities that govern the statistical complexity of the instance. $\delta_{min} := \min_{i\in[K], j\in[\Upsilon]} \delta_i^{(j)}$ is the minimum change between any two breakpoints, $\Delta_{min}^{(j)} := \min_{i\in[K], \Delta_i^{(j)}>0} \Delta_i^{(j)}$ and $\Delta_{max}^{(j)} := \max_{i\in[K]} \Delta_i^{(j)}$ are the minimum and maximum sub-optimality gap during an epoch $j \in [\Upsilon]$, respectively. Intuitively, the smaller $\delta_{min}^{(j)}$ is, the more difficult it is to detect breakpoints, and the smaller $\Delta_{min}^{(j)}$ is, the more difficult it is to distinguish the best action. On the other hand, the larger these quantities are, the larger the regret potentially incurred with an error. When (j) is omitted, we refer to the quantity minimized/maximized over all epochs.

2.3 Lower Bound

In this section, we provide a lower bound to the expected cumulative regret that any policy π must incur in an HTPS bandit.

Theorem 2.5 (Regret Lower Bound for the HTPS Bandit Problem)**.** *For any fixed policy π, we have*

$$\sup_{\boldsymbol{\nu}\in\mathcal{B}_{(v,\epsilon,\Upsilon)}} \mathbb{E}_{\boldsymbol{\nu}}[R^{\pi}(T)] \geq \frac{1}{25}(K\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}. \qquad (2)$$

Results of this type are known as *minimax lower bounds*. Indeed, the result states that, for every policy, there exists at least one instance in which the expected regret grows at a certain rate. The bound is consistent with the known lower bounds for the HT and PS MAB problems. Indeed, in HT MABs every policy has its expected regret lower bounded by $\Omega(K^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}})$ (Bubeck et al., 2013), while in PS MABs the lower bound is $\Omega(\sqrt{K\Upsilon T})$ (Garivier and Moulines, 2011). Thus, Equation (2) is a natural combination of these two results that can be recovered by either setting $\Upsilon = 1$ or $\epsilon = 1$, respectively. We refer to Appendix A for the proof.

3 TECHNICAL PRELIMINARIES

In this section, we introduce the technical tools we employ in our proposed solution. First, we discuss the mean estimation for HT random variables and describe the *Catoni estimator*. Then, we formalize the *change-point detection* (CPD) problem and discuss a technique based on *confidence sequences*.

3.1 Mean Estimation for Heavy-Tailed Random Variables with Catoni Estimator

Mean estimation for HT variables can be quite a delicate task. Empirical mean has been proven to achieve sub-optimal concentration (Bubeck et al., 2013). However, alternative estimators enjoying optimal rates have been proposed. We focus on the *Catoni estimator* (Catoni, 2012), defined using a *Catoni-type influence function* $\phi_\epsilon : \mathbb{R} \to \mathbb{R}$. The Catoni estimator $\widehat{\mu}_c$ for a sequence of variables $\{X_i\}_{i=1}^n$ is the solution of:

$$\sum_{i=1}^{n} \phi_\epsilon(\lambda_i(X_i - \widehat{\mu}_c)) = 0, \qquad (3)$$

where $\phi_\epsilon(x) = \log\left(1 + |x| + \frac{|x|^{1+\epsilon}}{1+\epsilon}\right)$, and $\{\lambda_i\}_{i=1}^n$ is a predictable process. Remarkably, for a proper choice of $\{\lambda_i\}_{i=1}^n$, this estimator enjoys an optimal concentration of order $\mathcal{O}\left(\left(\frac{v^{\frac{1}{\epsilon}} \log(\delta^{-1})}{n}\right)^{\frac{\epsilon}{1+\epsilon}}\right)$ with probability $1-\delta$ (Bhatt et al., 2022).

3.2 Confidence Sequences and Change Point Detection

The PS MAB is often addressed by resorting to *change-point detection* (CPD) strategies, *e.g.*, `CUSUM-UCB` (Liu et al., 2018). The idea is to actively adapt to environmental changes and tackle the problem as a sequence of stationary MABs. These strategies are often restricted to subgaussian rewards and do not scale on heavy-tailed variables. We propose an alternative approach to tackle this family of problems using a CPD strategy based on *confidence sequences*.

Confidence Sequences. Suppose $\{X_t\}_{t\in\mathbb{N}} \sim P$ for some $P \in \mathcal{P}^{\mu}$ where $\mathcal{P}^{\mu}$ is the set of distributions on $\prod_{t\in\mathbb{N}} \mathbb{R}$ such that $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = \mu$ for each $t \in \mathbb{N}$, where $\mathcal{F}_{t-1}$ is the filtration. A *confidence sequence* (CS) for the mean is a sequence of confidence intervals

$\{\text{CI}_t\}_{t\in\mathbb{N}}$ holding at arbitrary data-dependent stopping times. Formally:

$$\mathbb{P}(\forall t \in \mathbb{N}^+ : \mu \in \text{CI}_t) \geq 1-\gamma. \tag{4}$$

The random intervals $\{\text{CI}_t\}_{t\in\mathbb{N}^+}$ that satisfy property (4) are called $(1-\gamma)$-CS, where $1-\gamma$ is the confidence level. For a CS defined on $\mathbb{R}$, we formally introduce its *width* after t samples defined as $w(t,P,\gamma) \coloneqq \sup_{\mu_1,\mu_2\in CI_t} |\mu_1-\mu_2| \leq w(t,P,\gamma)$, for all $P \in \mathcal{P}^\mu$ and $\gamma \in (0,1]$.

Change-Point Detection. Consider a data-generating process composed of infinitely-countable distributions $\{P_t\}_{t\in\mathbb{N}}$ and let $t_c \geq 1$ be an unknown breakpoint, *i.e.*, $P_t = Q_0$ for every $t \leq t_c$ and $P_t = Q_1$ for every $t > t_c$. The goal of a CPD algorithm is to detect, as soon as possible after t_c, that a change in the data-generating distribution happened. In other words, given the (stochastic) stopping time $\tau \in \mathbb{N}$ in which the CPD system detects a change, the objective is to minimize the *detection delay* $\mathbb{E}_{t_c}[\tau - t_c]$, where the expectation $\mathbb{E}_{t_c}$ is taken over an environment having a change-point after t_c rounds. A trivial CPD system yielding a signal at every round would minimize this quantity. On the other hand, we also desire to reduce the *false alarm rate* (FAR), *i.e.*, the probability that a signal is produced when no change happens. This translates into minimizing $\mathbb{P}_\infty(\tau < \infty)$, where $\mathbb{P}_\infty$ is the probability measure of the environment where there is no change-point. Moreover, the *average run length* (ARL), defined as $\mathbb{E}_\infty[\tau]$, represents the expected number of rounds before a change is erroneously detected. However, when ARL is too large (*e.g.*, the trivial CPD that never yields a signal), the system is too conservative, impacting the detection delay. This highlights a crucial trade-off between detection delay and ARL. Recent literature (Shekhar and Ramdas, 2023a,b) sheds light on the possibility of reducing CPD to a sequential estimation, *i.e.*, producing sequential testing via confidence sequences. We focus on the `repeated-FCS-detector` framework, introduced in Shekhar and Ramdas (2023b). `repeated-FCS-detector` is a meta-algorithm that requires a CS computation strategy as a black-box tool, defined as:

Definition 3.1 (`repeated-FCS-detector`, Shekhar and Ramdas (2023b))**.** Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence of observations. At every round t, we receive a new sample X_t and initialize a new $(1-\gamma)$-CS for the mean $\text{CS}^{(t)} \coloneqq \{\text{CI}_n^{(t)}\}_{n\geq t}$, formed using samples X_t, X_{t+1}, X_{t+2}, and onwards. Moreover, we update all previously initialized CS $\{\text{CI}^{(i)}\}_{i<t}$ using X_t. We define the stopping time, τ, as the first time at which the intersection of all initialized CS becomes empty, *i.e.*, $\tau = \inf_{t\in\mathbb{N}}\{\bigcap_{n=0}^{t} \text{CS}^{(n)} = \emptyset\}$.

Provided an oracle capable of computing a $(1-\gamma)$-CS at every round, this strategy is distribution agnostic, as it requires no additional information about the data-generating distribution nor the change point. In Shekhar and Ramdas (2023b), theoretical guarantees on both the ARL and the detection delay of `repeated-FCS-detector` are provided in terms of the width of the $(1-\gamma)$-CS provided to the detector. We now report the theoretical guarantees.

Theorem 3.2 (Guarantees of `repeated-FCS-detector`, Shekhar and Ramdas (2023b))**.** *Consider a CPD problem with observations $\{X_t\}_{t\in\mathbb{N}}$ i.i.d. from $Q_0 \in \mathcal{P}^{\mu_0}$ for $t \leq t_c$ and from $Q_1 \in \mathcal{P}^{\mu_1}$ for $t > t_c$. Let $\delta \coloneqq |\mu_1 - \mu_0|$. Suppose we construct $(1-\gamma)$-confidence sequences with pointwise width $w(t,Q_0,\gamma)$ and $w(t,Q_1,\gamma)$ for pre- and post-change mean, respectively. Then, we have:* ***(i)*** *When there is no changepoint, the* `repeated-FCS-detector` *satisfies $\mathbb{E}_\infty[\tau] \geq \frac{1}{\gamma}$;* ***(ii)*** *Suppose $t_c < \infty$ and large enough to ensure that $w(t_c,\mu_0,\gamma) < \delta$. Define the event $\mathcal{E} \coloneqq \{\mu_0 \in \bigcap_{t=1}^{t_c} CI_t^{(1)}\}$, and note that $\mathbb{P}(\mathcal{E}) \geq 1-\gamma$ by construction. Then, for $\gamma \in (0,0.5)$, we have $\mathbb{E}_{t_c}[(\tau - t_c)^+ | \mathcal{E}] \leq \frac{3u_0(\mu_0,\mu_1,t_c)}{1-\gamma}$, where $u_0(\mu_0,\mu_1,t_c) \coloneqq \min_{n\in\mathbb{N}}\{w(n,\mu_1,\gamma) + w(t_c,\mu_0,\gamma) < \delta\}$.*

Point (i) provides a lower bound on the ARL of `repeated-FCS-detector`, while (ii) upper bounds the expected detection delay. While ARL only depends on the desired confidence level, the detection delay depends on the width of the CSs. Indeed, the width of the CS must decrease fast enough to make the change detectable. This assumption is standard in CPD, as enough samples before the change are needed to model the null hypothesis correctly.

4 ROBUST REGRET MINIMIZATION IN PIECEWISE-STATIONARY HEAVY-TAILED BANDITS

In this section, we describe our strategy for regret minimization in HTPS MABs. We start by providing a CPD strategy suited for heavy-tailed random variables together with its theoretical guarantees. Then, we leverage it to build a meta-algorithm named `Robust-CPD-UCB`, which uses a regret minimizer for the stationary setting and the CPD strategy to tackle non-stationarity.

4.1 `Catoni-FCS-detector`

We start by introducing a novel CPD strategy for HT random variables, which we name `Catoni-FCS-detector`, based on a `repeated-FCS-detector` using a special type of CS. We can define `Catoni-FCS-detector` as a special instantiation of `repeated-FCS-`

detector.

Definition 4.1 (Catoni-FCS-detector)**.** An instance of repeated-FCS-detector is a Catoni-FCS-detector if the $(1-\gamma)$-CS $\{\mathrm{CI}_t^{\phi}\}_{t\in\mathbb{N}}$ is defined as:

$$CI_t^{\phi} = \left\{ m \in \mathbb{R} : -\frac{v\sum_{i=1}^{t}\lambda_i^{1+\epsilon}}{2} - \log\left(\frac{2}{\gamma}\right) \right. \tag{5}$$

$$\left. \leq \sum_{i=1}^{t}\phi_\epsilon(\lambda_i(X_i - m)) \leq \frac{v\sum_{i=1}^{t}\lambda_i^{1+\epsilon}}{2} + \log\left(\frac{2}{\gamma}\right) \right\},$$

where ϕ_ϵ is the Catoni-type influence function (Equation 3).

From now on, we call *Catoni CS* the confidence sequence defined as in Equation (5). Catoni CS have been introduced for the first time in Wang and Ramdas (2023). While a Catoni CS does not admit a trivial closed-form representation, it can be proven (see Appendix A) that Equation (5) represents a proper $(1-\gamma)$-CS for the mean, attaining an optimal width (Bhatt et al., 2022). To use Catoni-FCS-detector in a bandit problem, however, we need specific types of guarantees, different than the ones provided for the general repeated-FCS-detector framework. We now provide two novel contributions of independent interest. First, we show how the width of the Catoni CS can be narrowed further w.r.t. to the one presented in previous works in the case of infinite variance. Second, we provide a finite-time bound on the detection delay of Catoni-FCS-detector, a crucial property for using a CPD in a bandit.

Proposition 4.2 (Detection Delay of Catoni-FCS-detector)**.** *Consider a CPD problem with observations $\{X_t\}_{t\in\mathbb{N}}$ drawn i.i.d. from $Q_0 \in \mathcal{H}_{\epsilon,v} \cap \mathcal{P}^{\mu_0}$ for $t \leq t_c$ and from $Q_1 \in \mathcal{H}_{\epsilon,v} \cap \mathcal{P}^{\mu_1}$ for $t > t_c$. Let $\delta := |\mu_1 - \mu_0|$. Suppose that there exists a known upper bound T of the change point ($t_c \leq T$). Let $n_{min} := 68\log(T^{\frac{1+\epsilon}{\epsilon}})$ and suppose $t_c \geq n_{min}$ large enough s.t. $w(t_c, Q_0, \gamma) \leq \frac{\delta}{2}$. Set $\gamma = \frac{2}{T^3}$. Then, there exists a predictable sequence $\{\lambda_i\}_{i=1}^{T}$ s.t.* Catoni-FCS-detector *enjoys **(i)** $\mathbb{P}_{t_c}\left((\tau - t_c)^+ \leq \mathcal{O}\left(v^{\frac{1}{\epsilon}}\frac{\log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}}\right)\right) \geq 1 - \frac{14}{T}$ and **(ii)** $\mathbb{P}_{t_c}(\tau < t_c) \leq \frac{14}{T}$.*

We point out the importance of this specialized result. Since the guarantees of Theorem 3.2 are very general, this result is aimed at providing a finite-time, high-probability bound on the detection delay when using Catoni CS. In particular, we make the term $u_0(\mu_0, \mu_1, t_c)$ from Theorem 3.2 explicit by using the properties of Catoni CS. Due to space reasons, the proof is postponed to Appendix A. Note that the rate of this detection delay cannot be improved as the lower bound for the detection delay of any distribution change is $\Omega(\log(\gamma^{-1}))$ (Lorden, 1971), where γ the confidence parameter that we set to $\mathcal{O}(1/T)$. Moreover, the dependencies on δ, ϵ and v may also be tight, as they embed the log-likelihood ratio of the test for the means of heavy-tailed random variables. We leave the answer to this question for further investigations. We conclude this section with two important remarks.

Algorithm 1: Robust-CPD-UCB

Input : Number of actions K, time horizon T, uniform exploration η, a policy π_s.
Initialize $t \leftarrow 0$, $t' \leftarrow 0$, $N_{i,t} \leftarrow 0 \quad \forall i \in [K]$.
Set $\gamma \leftarrow \frac{2}{T^3}$.
for $t \in [T]$ **do**
  if t' $mod \lfloor K/\eta \rfloor \leq K$ **then**
    Select and play $I_t \leftarrow t' \mod \lfloor K/\eta \rfloor$.
  end
  else
    Update π_s with the history of the last t' rounds.
    Select and play I_t according to π_s.
  end
  Receive X_t and update $N_{I_t,t} \leftarrow N_{I_t,t} + 1$ and $t' \leftarrow t' + 1$.
  if $N_{I_t,t} \geq n_{min}$ **then**
    Start a new $(1-\gamma)$-CS $\mathrm{CS}_{I_t}^{(t')}$ for action I_t, according to Equation (5).
    if $\exists a, b \in [t'] : CS_{I_t}^{(a)} \cap CS_{I_t}^{(b)} = \emptyset$ **then**
      Reset $t' \leftarrow 0$, $N_{i,t} \leftarrow 0 \quad \forall i \in [K]$.
      Remove all initialized CS.
    end
  end
end

Remark 4.3 (Comparison with Existing CPDs). Catoni-FCS-detector is, to the best of authors' knowledge, the first CPD strategy for the mean of HT random variables with infinite variance enjoying such guarantees. Thus, we consider our analysis an interesting standalone contribution. In bandit literature, however, many CPD strategies have been employed, (*e.g.*, CUSUM (Liu et al., 2018) and GLR Test (Besson et al., 2022)). However, they do not cover the HT scenario and often rely on strong parametric assumptions on the sample-generating distribution, *e.g.*, only working on Bernoulli variables.

Remark 4.4 (On the a priori knowledge of Catoni-FCS-detector). Catoni-FCS-detector does not rely, in principle, on any prior knowledge of the magnitude of the change or on the means. The confidence parameter γ is set based on the time horizon T, which is standard in MABs. Moreover, the sequence $\{\lambda_i\}_{i=1}^{t}$ can be set in advance for every $t \in [T]$, only relying on the knowledge of T.

4.2 Robust-CPD-UCB

In this section, we present Robust-CPD-UCB (R-CPD-UCB, Algorithm 1), an algorithm for PS HT bandits that adapts to environmental changes. It consists of three parts: (1) a sub-algorithm suited for the sta-

tionary HT MAB problem, that aims to minimize the regret in the stationary segments, we call this policy π_s; (2) the `Catoni-FCS-detector` strategy for CPD; and (3) a cyclic uniform exploration that ensures the availability of enough samples for every action to perform the CPD test. Algorithm 1 proceeds as follows: every $\lfloor K/\eta \rfloor$ rounds, it tries all the actions once (lines 4-5), which ensures that CPD can happen efficiently even when underrepresented actions in the history are the only ones changing. In the other rounds, a bandit sub-routine (*e.g.*, `Robust-UCB`) plays according to all the history since the last reset (lines 8-9); once the new reward is obtained, it is fed to the `Catoni-FCS-detector` that verifies if a change occurred (lines 12-14), in this case, everything is reset (line 15-16).

Remark 4.5. (Connection to `Monitored-UCB` from Cao et al. (2019)) `Robust-CPD-UCB` borrows the idea of cyclic uniform exploration from the `Monitored-UCB` (Cao et al., 2019). Like most of the algorithms for the PS setting, ours shares the usage of a stationary bandit sub-routine. However, a crucial difference lies in the type of CPD strategy employed. `Monitored-UCB` leverages a sliding-window type of CPD strategy that checks if the average of the first half of the sliding window is significantly different from that of the second half. This type of CPD strategy requires two hyperparameters, the window size and the threshold, respectively. Tuning these parameters may be difficult, even though, in practice, the algorithm works well even under misspecification. Finally, `Monitored-UCB` only deals with rewards bounded in $[0, 1]$, while `Robust-CPD-UCB` deals with HT rewards.

Remark 4.6. (Priori Knowledge of `R-CPD-UCB`) Algorithm 1 receives as inputs the time horizon T, the uniform exploration coefficient η, and a regret minimizer for the stationary setting π_s only. Assuming that it is possible to choose a regret minimizer that does not require additional parameters other than T (which is, as we will show in the next section, rather natural), then the only knowledge that `R-CPD-UCB` requires on the environment is the time horizon T. In practice, this property ensures that no tuning must happen.

4.3 Theoretical Guarantees of `Robust-CPD-UCB`

As customary in the literature of PS MABs, we introduce a technical assumption regarding the length of any epoch, ensuring that exploration is frequent enough to detect for every action.

Assumption 4.7. For every epoch $j \in [\Upsilon]$, let $\widetilde{\delta}^{(j)}_{min} := \min\{\delta^{(j-1)}_{min}, \delta^{(j)}_{min}\}$, and let $|E_j|$ be its length and $L_j := 6(236)^{\frac{1+\epsilon}{\epsilon}} v^{\frac{1}{\epsilon}} \frac{\log\left(\log\left(1/\widetilde{\delta}^{(j)}_{min}\right)\right)+\log(T)}{(\widetilde{\delta}^{(j)}_{min})^{\frac{1+\epsilon}{\epsilon}}}$. Let $n_{min} := 68\log(T^{\frac{1+\epsilon}{\epsilon}})$. Then, the learner can select η such that, for every $j \in [\Upsilon]$, it holds that $|E_j| \geq 2n_{min} + 2\lceil L_j K/\eta \rceil$.

This ensures that proper learning can be performed in such an environment Indeed, we enforce that every epoch $j \in [\Upsilon]$ is large enough so that, due to the forced exploration only, the algorithm chooses every action at least L_j times. This kind of assumption is ubiquitous in the piecewise-stationary bandits literature. Notable examples include Assumptions 4 and 7 in Besson et al. (2022), Assumptions 1 and 2 in Cao et al. (2019) and Assumption 1 in Liu et al. (2018). Some are equivalent to ours, while others are neither weaker nor stronger. Alternative assumptions, such as the monotonicity of the mean change, also allow for theoretical tractability, *e.g.*, Assumption 1 in Seznec et al. (2020), which forces expected rewards to evolve in a decreasing manner. Note that Algorithm 1 can operate regardless of this assumption, as shown in Section 5. We are now ready to present our main result.

Theorem 4.8 (Regret Upper Bound of `R-CPD-UCB`)**.** *Under Assumption 4.7,* `R-CPD-UCB` *suffers an expected cumulative regret bounded as:*

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \mathcal{O}\Bigg(\underbrace{v^{\frac{1}{\epsilon}} \sum_{j=1}^{\Upsilon} \frac{\log(T)}{\left(\widetilde{\delta}^{(j)}_{min}\right)^{\frac{1+\epsilon}{\epsilon}}} \left\lceil \frac{K}{\eta} \right\rceil \Delta^{(j)}_{max}}_{\texttt{(A)}\text{ Detection Delay Contribution}} + \underbrace{\sum_{j=1}^{\Upsilon} \mathbb{E}[R^{\pi_s}(|E_j|)]}_{\texttt{(B)}\text{ Stationary Policy Regret}} + \underbrace{\eta \sum_{j=1}^{\Upsilon} |E_j| \Delta^{(j)}_{max}}_{\texttt{(C)}\text{ Uniform Exploration}} \Bigg). \quad (6)$$

The regret can be decomposed into three contributions due to: the detection delay (part (A)), the regret-per-epoch of the stationary policy (part (B)), and the rounds of uniform exploration (part (C)).

Uniform Exploration Trade-off. The uniform exploration parameter η appears in both (A) and (C). Setting aside part (B), it is clear that η creates a trade-off between these two: the larger η is, the quicker the algorithm can detect a change, and the smaller is (A); on the other hand, excessive uniform exploration inflates the regret of `R-CPD-UCB` and the contribution from (C). Finding the optimal value for η would require extensive prior knowledge, which is, in general, not available. A good trade-off is to set $\eta = \sqrt{\Upsilon/T}$, which impose both (A) and (C) to be $\widetilde{\mathcal{O}}(\sqrt{\Upsilon T})$. However, it is possible to define a forced exploration strategy that does not require any knowledge of Υ, making the algorithm more versatile while keeping the same order of performance. In particular, we leverage the methodology developed in Besson et al. (2022) and obtain the following result.

Corollary 4.9. *Let* $\{\eta_j\}_{j\in\mathbb{N}}$ *where* $\eta_j =$

$\eta_0\sqrt{jK\log(T)/T}$ for some $\eta_0 > 0$ be an increasing sequence. `R-CPD-UCB` *using η_{j+1} after the j-th detection satisfies:*

$$(A) \leq \frac{v^{\frac{1}{\epsilon}}\sqrt{K\Upsilon T\log(T)}}{\eta_0 \delta_{min}^{\frac{1+\epsilon}{\epsilon}}}\Delta_{max}, \tag{7}$$

$$(C) \leq \eta_0\sqrt{K(\Upsilon+1)T\log(T)}\Delta_{max}. \tag{8}$$

Note that, if δ_{min} is known, setting $\eta_0 = \delta_{min}^{-\frac{1+\epsilon}{2\epsilon}}$ can further reduce the regret bound.

Choosing π_s. The choice of the inner regret minimizer π_s determines the magnitude of part (B). The best choice is to select a policy that has a regret upper bound matching the known lower bound of $\Omega(K^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}})$. We can instantiate `R-CPD-UCB` using the `Robust UCB` policy with *median-of-means estimator* from (Bubeck et al., 2013, Section 2.2). As a result, we get the following bounds.

Corollary 4.10. *Let π_s be the* `Robust UCB` *policy with median-of-means estimator (Bubeck et al., 2013, Section 2.2). Under Assumption 4.7,* `R-CPD-UCB` *suffers an expected cumulative regret bounded as:*

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \mathcal{O}\Bigg((A) + \underbrace{\sum_{j=1}^{\Upsilon}\sum_{i:\Delta_i^{(j)}>0}\frac{v^{\frac{1}{\epsilon}}\log\left(|E_j|\right)}{\left(\Delta_i^{(j)}\right)^{\frac{1}{\epsilon}}}}_{\substack{(B_1)\text{ Robust UCB Regret}\\ \text{(Instance Dependent)}}} + (C)\Bigg). \tag{9}$$

Moreover, if $\log(|E_j|) \geq \frac{5(\Delta_{max}^{(j)})^{\frac{1+\epsilon}{\epsilon}}}{2v^{\frac{1}{\epsilon}}}$ for every $j \in [\Upsilon]$, we have

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \widetilde{\mathcal{O}}\Bigg((A) + \underbrace{(K\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}}_{\substack{(B_2)\text{ Robust UCB Regret}\\ \text{(Instance Independent)}}} + (C)\Bigg). \tag{10}$$

Equation (9) is a direct consequence of Theorem 4.8 and Theorem 3 of Bubeck et al. (2013). Equation (10) follows from Theorem 4.8, Proposition 1 of Bubeck et al. (2013), and Jensen's inequality. `Robust UCB` enjoys both instance-dependent and instance-independent guarantees: part (B_1) depends on the sub-optimality gaps $\Delta_i^{(j)}$ and the individual lengths of the epochs, while part (B_2) does not, as it accounts for a worst-case scenario of the sub-optimality gaps. We can now combine all and get the following.

Corollary 4.11. *Let π_s be the* `Robust UCB` *policy with median-of-means estimator from (Bubeck et al., 2013, Section 2.2). Let $\{\eta_j\}_{j\in\mathbb{N}}$ where $\eta_j = \eta_0\sqrt{jK\log(T)/T}$ for some $\eta_0 > 0$. Under Assumption 4.7,* `R-CPD-UCB` *using η_{j+1} after the j-th detection suffers an expected cumulative regret bounded as:*

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \mathcal{O}\left(\frac{v^{\frac{1}{\epsilon}}\sqrt{K\Upsilon T\log(T)}}{\eta_0\delta_{min}^{\frac{1+\epsilon}{\epsilon}}}\Delta_{max} + \frac{K\Upsilon v^{\frac{1}{\epsilon}}\log(T/\Upsilon)}{\Delta_{min}^{\frac{1}{\epsilon}}}\right). \tag{11}$$

Moreover, if $\log(|E_j|) \geq 3(\Delta_{max}^{(j)})^{\frac{1+\epsilon}{\epsilon}}v^{-\frac{1}{\epsilon}}$ for every $j \in [\Upsilon]$, and $\delta_{min}^{\frac{1+\epsilon}{\epsilon}} \geq v^{\frac{1}{\epsilon(1+\epsilon)}}(\Upsilon K/T)^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\log(T)}$, we have

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \widetilde{\mathcal{O}}\Big((K\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}\Big). \tag{12}$$

Equation (11) depends on both the minimum mean change δ_{min}, and the extreme sub-optimality gaps Δ_{min} and Δ_{max}, along the whole trial. We consider this bound an instance-dependent guarantee over the performance of `R-CPD-UCB`. Equation (12), instead, does not contain any of these quantities. The second assumption fundamentally states that δ_{min} can be assumed to be a constant w.r.t. the other quantities, in particular T. In this case, an instance-independent bound can be obtained. Equation (12) matches, up to constants, the lower bound presented in Theorem 2.5. Thus, the performance guarantees of `R-CPD-UCB` are nearly-optimal w.r.t. T, Υ, v, and K

5 NUMERICAL EVALUATION

We now provide a numerical evaluation of `Robust-CPD-UCB` (π_s chosen as `Robust UCB` with median-of-means estimator). We refer to Appendix C for additional details and experimental campaigns.

5.1 Casting Real-World Data to HTPS MABs

We model a real-world scenario as an HTPS MAB and use a real dataset to build an instance on which `R-CPD-UCB` is tested.

Setting. We study profit maximization in financial trading, where data are heavy-tailed (Panahi, 2016). A natural application is selecting the most profitable cryptocurrency among K options: each day, an investor invests in the currency with the highest closing price. This problem has been explored in HT MABs by Yu et al. (2018); Lee and Lim (2022). We use the same dataset (Kaggle link) as Lee and Lim (2022). Figure 1 shows closing prices of four top-10 cryptocurrencies with a piecewise-constant fit minimizing squared error. Two observations follow: (i) this fit outperforms any constant approximation (where rewards were assumed stationary in previous HT MAB works); (ii) strong

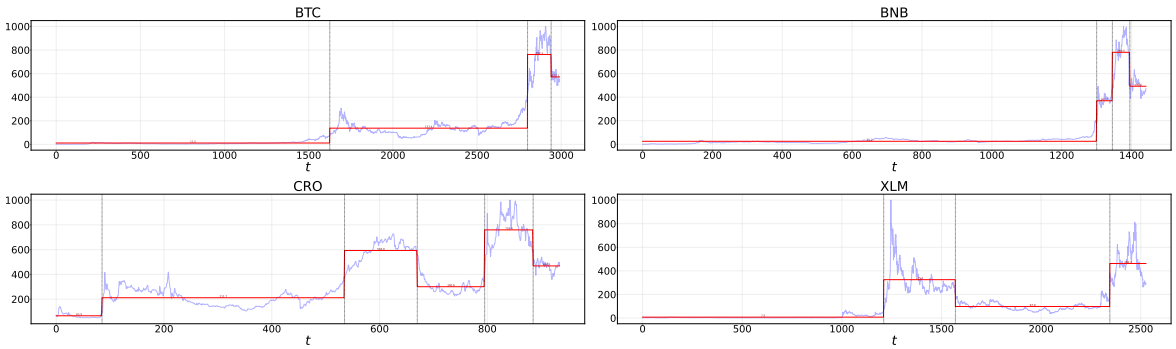


Figure 1: Rescaled closing prices of four selected cryptocurrencies (blue) with a piecewise-constant approximation (red). Each time step is a day starting in April 2016. Source: Kaggle Dataset.

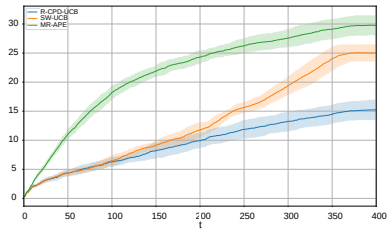
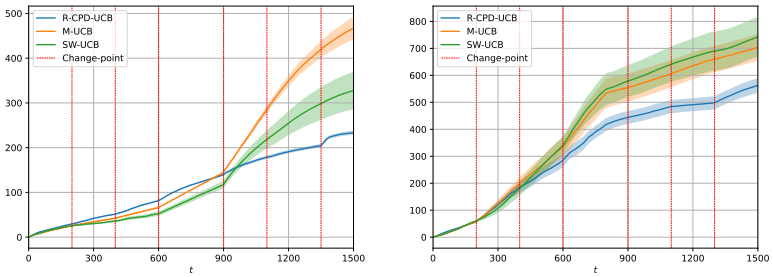


Figure 2: Cumulative regrets on HTPS built from cryptocurrency dataset. 20 trials, mean ± std.



(a) Gaussian rewards. (b) Pareto rewards.

Figure 3: Cumulative regrets. 20 trials, mean ± std.

fluctuations yield high errors in certain segments. Following Panahi (2016), we model each segment with a Pareto distribution centered on the segment mean, with $\epsilon < 1$ and $v = 3$. Thus, cryptocurrency profit maximization is an HTPS MAB. From the piecewise-constant fit, we construct an HTPS MAB environment and test `R-CPD-UCB` against `Sliding Window UCB` (Garivier and Moulines, 2011) and `MR-APE` (Lee and Lim, 2022), the latter already evaluated on this dataset under stationarity.

Results. Figure 2 reports cumulative regrets averaged over 20 trials (rescaled x-axis). `R-CPD-UCB` outperforms both, being the only algorithm handling heavy-tailedness and non-stationarity.

5.2 Regret Minimization in Highly Non-Stationary Environments

We now evaluate `R-CPD-UCB` in highly dynamic scenarios with closely spaced changes.

Setting. We compare `R-CPD-UCB` with two baselines, `Monitored UCB` (Cao et al., 2019) and `Sliding Window UCB` (Garivier and Moulines, 2011), on two PS MABs: Gaussian rewards with $\sigma = 1$ and Pareto rewards with $\epsilon < \frac{1}{2}$. In both cases, $K = 3$, $T = 1500$, and $\Upsilon = 6$. We set $\delta_{min} = 0$, so some actions may not change after a change point, but at least one arm always does, with the optimal arm switching at least four times. These instances violate Assumption 4.7.

Results. Figure 5 shows cumulative regrets. In both settings, `R-CPD-UCB` achieves lower regret than competitors, with smaller uncertainty and more stable performance, especially with infinite variance (Figure 5b). Remarkably, `R-CPD-UCB` also outperforms both baselines on Gaussian rewards, due to frequent, closely spaced change points where median-of-means stabilizes learning in data-scarce regimes. Although Assumption 4.7 is violated, `R-CPD-UCB` (and `Monitored UCB`) still perform well, confirming observations from Cao et al. (2019) that this assumption is not restrictive in practice.

6 CONCLUSIONS

We presented the first study of regret minimization in heavy-tailed piecewise-stationary bandits, establishing a lower bound and proposing `Robust-CPD-UCB`, whose regret nearly matches it. Leveraging recent advances in change-point detection, we introduced `Catoni-FCS-detector`, a strategy for distributions with infinite variance. Numerical experiments show that `R-CPD-UCB` outperforms baselines from the literature. A future direction is to study HTPS MABs when v and ϵ are unknown (Genalti and Metelli, 2025).

Acknowledgements

This publication was funded with the contribution of Ministero dell'Università e della ricerca pursuant to D.D. n. 7206 of 17 April 2025 – BANDO FIS 2. Project FIS-2023-02598 (Starting Grant), title: "Unified Learning from Diverse Human Feedback" (HUmLrn). CUP: D53C25000710001

References

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

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

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ϵ, u)-adaptive regret minimization in heavy-tailed bandits. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1882–1915. PMLR, 2024a.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th annual international conference on machine learning*, pages 1177–1184, 2009.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International conference on algorithmic learning theory*, pages 174–188. Springer, 2011.

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.

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

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential changepoint detection via backward confidence sequences. In *International Conference on Machine Learning*, pages 30908–30930. PMLR, 2023a.

Shubhanshu Shekhar and Aaditya Ramdas. Reducing sequential change detection to sequential estimation. *arXiv preprint arXiv:2309.09111*, 2023b.

Hongjian Wang and Aaditya Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and Their Applications*, 163:168–202, 2023.

Gary Lorden. Procedures for reacting to a change in distribution. *The annals of mathematical statistics*, pages 1897–1908, 1971.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(77):1–40, 2022.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 418–427. PMLR, 2019.

Julien Seznec, Pierre Menard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3784–3794. PMLR, 2020.

Hanieh Panahi. Model selection test for the heavy-tailed distributions under censored samples with application in financial data. *International Journal of Financial Studies*, 4(4):24, 2016.

Xiaotian Yu, Han Shao, Michael R Lyu, and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *UAI*, pages 937–946, 2018.

Kyungjae Lee and Sungbin Lim. Minimax optimal bandits for heavy tail rewards. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5280–5294, 2022.

Gianmarco Genalti and Alberto Maria Metelli. Open problem: Regret minimization in heavy-tailed bandits with unknown distributional parameters. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 1–5. PMLR, 2025.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.

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

Cédric Hartland, Nicolas Baskiotis, Sylvain Gelly, Michele Sebag, and Olivier Teytaud. Change point detection and meta-bandits for online learning in dynamic environments. In *CAp 2007: 9è Conférence francophone sur l'apprentissage automatique*, pages 237–250, 2007.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158. PMLR, 2019.

Sujay Bhatt, Guanhua Fang, and Ping Li. Piecewise stationary bandits under risk criteria. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4335. PMLR, 2023.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2564–2572. PMLR, 2019.

Alberto Maria Metelli, Francesco Trovo, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *International Conference on Machine Learning*, pages 15421–15457. PMLR, 2022.

Hoda Heidari, Michael J Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *IJCAI*, pages 1562–1570, 2016.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto Maria Metelli. Graph-triggered rising bandits. In *Forty-first International Conference on Machine Learning*, 2024b.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [No]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Proofs

Theorem 2.5 (Regret Lower Bound for the HTPS Bandit Problem)**.** *For any fixed policy π, we have*

$$\sup_{\boldsymbol{\nu}\in\mathcal{B}_{(v,\epsilon,\Upsilon)}} \mathbb{E}_{\boldsymbol{\nu}}[R^{\pi}(T)] \geq \frac{1}{25}(K\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}. \tag{2}$$

Proof. The proof of this theorem combines techniques from Lemma 5 of Seznec et al. (2020), Theorem 4 of Genalti et al. (2024a), and Theorem 6 from Garivier et al. (2019).

Consider the following prototype of reward distribution, defined for $y \in (0,1)$ and $\Delta \in (0,1)$:

$$\rho_y = \left(1 - v^{-\frac{1}{\epsilon}} y^{\frac{1+\epsilon}{\epsilon}}\right)\delta_0 + \left(v^{-\frac{1}{\epsilon}} y^{\frac{1+\epsilon}{\epsilon}}\right)\delta_{v^{\frac{1}{\epsilon}}\Delta^{-\frac{1}{\epsilon}}}.$$

It is easy to verify that $\rho_y \in \mathcal{H}_{(1,\epsilon)}$ for every $y \in [0,\Delta]$.

Consider a set of of instances belonging to $\mathcal{B}_{(v,\epsilon,\Upsilon)}$ indexed by a vector $\boldsymbol{i}^* \in [K]^{\Upsilon}$ in a way that, for every $j \in [\Upsilon]$ and every $t \in E_j$, we have

$$\nu_i^{(j)} = \begin{cases} \rho_{2^{\frac{\epsilon}{1+\epsilon}}\Delta}, & \text{if } i = i_j^* \\ \rho_{\Delta}, & \text{if } i \neq i_j^* \end{cases}.$$

It follows that $\mu_{i_j^*}^{(j)} - \mu_i^{(j)} = \Delta$ for every $j \in [\Upsilon]$ and $i \neq i_j^*$. Let $|E_j| = \frac{T}{\Upsilon}$ for every $j \in [\Upsilon]$, assuming w.l.o.g. that T is divisible by Υ. Thus, all epochs are of the same length. For every fixed policy π, we write the average expected regret among the instances indexed by $\boldsymbol{i}^*$:

$$\begin{aligned}
\frac{1}{K^{\Upsilon}}\sum_{\boldsymbol{i}^*\in[K]^{\Upsilon}} \mathbb{E}_{\boldsymbol{i}^*}\left[R^{\pi}(T)\right] &= \frac{1}{K^{\Upsilon}}\sum_{\boldsymbol{i}^*\in[K]^{\Upsilon}}\sum_{j=1}^{\Upsilon}\Delta\mathbb{E}_{\boldsymbol{i}^*}\left[|E_j| - N_{i_j^*}^{(j)}\right] \\
&= \Delta\left(T - \frac{1}{K^{\Upsilon}}\sum_{\boldsymbol{i}^*\in[K]^{\Upsilon}}\sum_{j=1}^{\Upsilon}\mathbb{E}_{\boldsymbol{i}^*}\left[N_{i_j^*}^{(j)}\right]\right) \\
&= \Delta\left(T - \sum_{j=1}^{\Upsilon}\frac{1}{K^{\Upsilon-1}}\sum_{\boldsymbol{i}_{-j}^*\in[K]^{\Upsilon-1}}\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{(\boldsymbol{i}_{-j}^*,i)}\left[N_i^{(j)}\right]\right),
\end{aligned} \tag{13}$$

where $\boldsymbol{i}_{-j}^*$ equals to $\boldsymbol{i}^*$ where the j-th coordinate is set to 0 and $(\boldsymbol{i}_{-j}^*, i)$ equals to $\boldsymbol{i}^*$ where the j-th coordinate is set to i, for $i \in [K]$.

Let $D_{KL}(P,Q)$ be the Kullback-Leibler divergence between P and Q, then we have:

$$\begin{aligned}
D_{KL}\left(\rho_{2^{\frac{\epsilon}{1+\epsilon}}\Delta}, \rho_{\Delta}\right) &= \left(1 - 2v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}\right)\log\left(\frac{1 - 2v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}}{1 - v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}}\right) + 2v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}\log\left(\frac{2v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}}{v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}}\right) \\
&\leq 2v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}\log(2),
\end{aligned}$$

where the inequality follows by upper bounding the first addendum with 0. Using Pinsker Inequality and the

previous bound on the KL divergence, for every $j \in [\Upsilon]$, we get

$$
\begin{aligned}
2\left(\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{(\boldsymbol{i}^*_{-j},i)}\left[\frac{N_i^{(j)}}{|E_j|}\right]-\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{(\boldsymbol{i}^*_{-j},0)}\left[\frac{N_i^{(j)}}{|E_j|}\right]\right)^2 &\leq D_{KL}\left(\mathbb{P}_{(\boldsymbol{i}^*_{-j},i)},\mathbb{P}_{(\boldsymbol{i}^*_{-j},0)}\right)\\
&\leq \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{(\boldsymbol{i}^*_{-j},i)}\left[N_i^{(j)}\right]D_{KL}\left(\rho_{2\frac{\epsilon}{1+\epsilon}\Delta},\rho_{\Delta}\right)\\
&\leq \frac{\log(2)}{2K}|E_j|v^{-\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}},
\end{aligned}
$$

that implies

$$
\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{(\boldsymbol{i}^*_{-j},i)}\left[N_i^{(j)}\right]\leq\frac{|E_j|}{K}+\sqrt{\frac{\log(2)}{2K}}|E_j|^{\frac{3}{2}}v^{-\frac{1}{2\epsilon}}\Delta^{\frac{1+\epsilon}{2\epsilon}}. \tag{14}
$$

Combining Equation (13) and Equation (14), we get

$$
\begin{aligned}
\frac{1}{K^{\Upsilon}}\sum_{\boldsymbol{i}^*\in[K]^{\Upsilon}}\mathbb{E}_{\boldsymbol{i}^*}\left[R^{\pi}(T)\right] &\geq \left(\frac{T}{2}-\sum_{k=1}^{\Upsilon}\sqrt{\frac{\log(2)}{2K}}|E_j|^{\frac{3}{2}}v^{-\frac{1}{2\epsilon}}\Delta^{\frac{1+\epsilon}{2\epsilon}}\right)\Delta\\
&\geq \frac{1}{2}\left(\frac{2\log(2)}{16}\right)^{\frac{\epsilon}{1+\epsilon}}K^{\frac{\epsilon}{1+\epsilon}}\Upsilon^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}},
\end{aligned}
$$

by setting $\Delta = v^{\frac{1}{1+\epsilon}}\left(\frac{2\log(2)K\Upsilon}{16T}\right)^{\frac{\epsilon}{1+\epsilon}}$. □

Remark A.1. In the proof of Theorem 2.5, we do not impose any condition on T. Moreover, the length of every epoch is equal to T/Υ. This means that in principle one can choose a large enough T, *i.e.*, $T \geq 2\Upsilon\lceil L_j\frac{K}{\eta}\rceil$, such that Assumption 4.7 holds. This proves that our assumption does not make the problem easier from a regret minimization perspective.

Proposition 4.2 (Detection Delay of `Catoni-FCS-detector`)**.** *Consider a CPD problem with observations $\{X_t\}_{t\in\mathbb{N}}$ drawn i.i.d. from $Q_0 \in \mathcal{H}_{\epsilon,v}\cap\mathcal{P}^{\mu_0}$ for $t \leq t_c$ and from $Q_1 \in \mathcal{H}_{\epsilon,v}\cap\mathcal{P}^{\mu_1}$ for $t > t_c$. Let $\delta := |\mu_1-\mu_0|$. Suppose that there exists a known upper bound T of the change point ($t_c \leq T$). Let $n_{min} := 68\log(T^{\frac{1+\epsilon}{\epsilon}})$ and suppose $t_c \geq n_{min}$ large enough s.t. $w(t_c, Q_0, \gamma) \leq \frac{\delta}{2}$. Set $\gamma = \frac{2}{T^3}$. Then, there exists a predictable sequence $\{\lambda_i\}_{i=1}^{T}$ s.t.* `Catoni-FCS-detector` *enjoys **(i)** $\mathbb{P}_{t_c}\left((\tau-t_c)^+ \leq \mathcal{O}\left(v^{\frac{1}{\epsilon}}\frac{\log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}}\right)\right) \geq 1-\frac{14}{T}$ and **(ii)** $\mathbb{P}_{t_c}(\tau < t_c) \leq \frac{14}{T}$.*

Proof. Due to its length, we divided this proof into several steps. In Steps 1-3 we extend Theorem 10 of Wang and Ramdas (2023) to the case of heavy-tailed random variables. In Step 4 we apply the *sticthing* technique to the resulting CS, tightening its width. Then, in Step 5 we define the CS hyper-parameters and define a set of good events, under which we are able to properly bound the detection delay in Step 7. The proof is concluded by showing that no false alarm occurs under the good event (Step 8).

Step 1 (Building a nonnegative supermartingale) First, we observe that

$$
\begin{aligned}
M_t &:= \prod_{i=1}^{t}\exp\left\{\phi_{\epsilon}(\lambda_i(X_i-\mu))-\lambda_i^{1+\epsilon}\frac{v}{1+\epsilon}\right\},\\
N_t &:= \prod_{i=1}^{t}\exp\left\{-\phi_{\epsilon}(\lambda_i(X_i-\mu))-\lambda_i^{1+\epsilon}\frac{v}{1+\epsilon}\right\},
\end{aligned}
$$

are nonnegative supermartingales. To prove this for M_t (all steps are analogous for N_t), we bound

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\phi_\epsilon(\lambda_t(X_t-\mu)) - \lambda_t \frac{v}{1+\epsilon}\right\} \,\middle|\, \mathcal{F}_{t-1}\right] \leq \\
&\leq \mathbb{E}\left[1+\lambda_t(X_t-\mu)+\lambda_t^{1+\epsilon}\frac{(X_t-\mu)^{1+\epsilon}}{1+\epsilon} \,\middle|\, \mathcal{F}_{t-1}\right]\exp\left\{-\lambda_t^{1+\epsilon}\frac{v}{1+\epsilon}\right\} \\
&\leq \left(1+\lambda_t^{1+\epsilon}\frac{v}{1+\epsilon}\right)\exp\left\{-\lambda_t^{1+\epsilon}\frac{v}{1+\epsilon}\right\} \leq 1,
\end{aligned}$$

and, subsequently,

$$\mathbb{E}\left[M_t \,\middle|\, \mathcal{F}_{t-1}\right] = M_{t-1}\mathbb{E}\left[\exp\left\{\phi_\epsilon(\lambda_t(X_t-\mu)) - \lambda_t^{1+\epsilon}\frac{v}{1+\epsilon}\right\} \,\middle|\, \mathcal{F}_{t-1}\right] \leq M_{t-1}.$$

Step 2 (Building a CS for ϕ_ϵ) Then, we can leverage Ville's inequality to construct a CS around $\phi_\epsilon(\lambda(X-\mu))$:

$$\mathbb{P}\left(\exists t \geq 1 : M_t \geq \frac{2}{\gamma}\right) \leq \frac{\gamma}{2},$$

which implies

$$\mathbb{P}\left(\exists t \geq 1 : \sum_{i=1}^t \phi_\epsilon(\lambda_i(X_i-\mu)) \geq \frac{v\sum_{i=1}^t \lambda_i^{1+\epsilon}}{1+\epsilon} + \log\left(\frac{2}{\gamma}\right)\right) \leq \frac{\gamma}{2}.$$

Analogous calculations for N_t and a union bound, yield a $(1-\gamma)$-CS where the intervals have the following form:

$$CI_t^\phi = \left\{m \in \mathbb{R} : -\frac{v\sum_{i=1}^t \lambda_i^{1+\epsilon}}{1+\epsilon} - \log\left(\frac{2}{\gamma}\right) \leq \sum_{i=1}^t \phi_\epsilon(\lambda_i(X_i-m)) \leq \frac{v\sum_{i=1}^t \lambda_i^{1+\epsilon}}{1+\epsilon} + \log\left(\frac{2}{\gamma}\right)\right\}.$$

Step 3 (Bounding the width of the CS for μ) We are now required to provide a bound on the width of the previously derived $(1-\gamma)$-CS. To do so, we derive high-probability lower and upper bounds over the random solution of $f_t(m) \coloneqq \sum_{i=1}^t \phi_\epsilon(\lambda_i(X_i-m)) = 0$. For all $m \in \mathbb{R}$, let

$$M_t(m) = \exp\left\{f_t(m) - \sum_{i=1}^t \left(\lambda_i(\mu-m) + \frac{\lambda_i^{1+\epsilon}}{1+\epsilon}(v+(\mu-m)^{1+\epsilon})\right)\right\},$$

then, with steps analogous to Step 1, we observe that $M_t(m)$ is a nonnegative supermartingale. Note that $M_t(\mu) = M_t$, and an analogous definition leads to the nonnegative supermartingale $N_t(m)$. We define:

$$\begin{aligned}
B_t^+(m) &= \sum_{i=1}^t \left(\lambda_i(\mu-m) + \frac{\lambda_i^{1+\epsilon}}{1+\epsilon}(v+(\mu-m)^{1+\epsilon})\right) + \log\left(\frac{2}{\theta}\right) \\
B_t^-(m) &= \sum_{i=1}^t \left(\lambda_i(\mu-m) - \frac{\lambda_i^{1+\epsilon}}{1+\epsilon}(v+(\mu-m)^{1+\epsilon})\right) - \log\left(\frac{2}{\theta}\right),
\end{aligned}$$

and using Markov's ineqality we get:

$$\begin{aligned}
\forall m \in \mathbb{R}, \quad \mathbb{P}\left(f_t(m) \leq B_t^+(m)\right) &\geq 1 - \frac{\theta}{2} \\
\forall m \in \mathbb{R}, \quad \mathbb{P}\left(f_t(m) \geq B_t^-(m)\right) &\geq 1 - \frac{\theta}{2}.
\end{aligned}$$

Since B_t^+ upper bounds $f_t(m)$ with probability at least $1-\frac{\theta}{2}$, any $\widetilde{m}_t$ s.t.

$$B_t^+(\widetilde{m}_t) = -\sum_{i=1}^t \frac{v\lambda_i^{1+\epsilon}}{1+\epsilon} - \log\left(\frac{2}{\gamma}\right) = f_t(\max\{CI_t^\phi\}) \tag{15}$$

also satisfies

$$\mathbb{P}\left(f_t(\widetilde{m}_t) \leq f_t(\max\{CI_t^{\phi}\}\right) \geq 1 - \frac{\theta}{2}, \tag{16}$$

where $\widetilde{m}_t$ is a non-random quantity as it's the solution to a deterministic equation. As $f_t(m)$ is a non-increasing function of m, Equation (16) implies that:

$$\mathbb{P}\left(\widetilde{m}_t \leq \max\{CI_t^{\phi}\}\right) \geq 1 - \frac{\theta}{2}.$$

Note that Equation (15) admits solutions if and only if

$$\left(\sum_{i=1}^{t} \lambda_i^{1+\epsilon}\right)^{\frac{1}{\epsilon}} \left(\sum_{i=1}^{t} \lambda_i\right)^{-\frac{1+\epsilon}{\epsilon}} \left(\sum_{i=1}^{t} 5\lambda_i^{1+\epsilon} \frac{v}{1+\epsilon} + 2\log\left(\frac{2}{\gamma}\right) + 2\log\left(\frac{2}{\theta}\right)\right) \leq \frac{\epsilon}{1+\epsilon}. \tag{17}$$

Finally, we conclude this step by bounding

$$\widetilde{m}_t \leq \mu + \frac{\sum_{i=1}^{t} 10v\lambda_i^{1+\epsilon} + 2(1+\epsilon)\log\left(\frac{2}{\gamma}\right) + 2(1+\epsilon)\log\left(\frac{2}{\theta}\right)}{\sum_{i=1}^{t} \lambda_i},$$

which yields the upper CS on μ in the following form:

$$\mathbb{P}\left(\max\{CI_t^{\phi}\} \leq \mu + \frac{\sum_{i=1}^{t} 10v\lambda_i^{1+\epsilon} + 2(1+\epsilon)\log\left(\frac{2}{\gamma}\right) + 2(1+\epsilon)\log\left(\frac{2}{\theta}\right)}{\sum_{i=1}^{t} \lambda_i}\right) \geq 1 - \frac{\theta}{2}.$$

Repeating all the previous steps for $B_t^{-}(m)$, and by applying a union bound, yields a two-sided $(1-\gamma)$-CS for μ. The width $w(t,\mu,\gamma) = \max\{CI_t^{\phi}\} - \min\{CI_t^{\phi}\}$ of such CS concentrates as

$$\mathbb{P}\left(w(t,\mu,\gamma) \leq 2\frac{\sum_{i=1}^{t} 10v\lambda_i^{1+\epsilon} + 2(1+\epsilon)\log\left(\frac{2}{\gamma}\right) + 2(1+\epsilon)\log\left(\frac{2}{\theta}\right)}{\sum_{i=1}^{t} \lambda_i}\right) \geq 1 - \theta. \tag{18}$$

Step 4 (Stitching) We now discuss the choice of the sequence $\{\lambda_t\}_{t\geq 1}$. The idea is to partition time in an exponential grid, and then fix the same value of λ_t inside the same cell. Moreover, the confidence level is modified and set to a cell-specific value γ_j. This idea, called *stitching*, first appeared in Howard et al. (2021). In particular, set $t_j = e^j$, $\gamma_j = \frac{\gamma}{(j+1)^2}$, and $\Lambda_j = \left(\log\left(\frac{2}{\gamma_j}\right) e^{-j} v^{-1}\right)^{\frac{1}{1+\epsilon}}$. Then, for every $t_j < t \leq t_{j+1}$, we set $\lambda_i = \Lambda_j$ for every $i \in [t]$. Assume $\theta = \frac{\gamma}{4}$. For every $t_j < t \leq t_{j+1}$, we have

$$\begin{aligned}
&\frac{\sum_{i=1}^{t} 10v\lambda_i^{1+\epsilon} + 2(1+\epsilon)\log\left(\frac{2}{\gamma_j}\right) + 2(1+\epsilon)\log\left(\frac{2}{\theta}\right)}{\sum_{i=1}^{t} \lambda_i} \\
&= v^{\frac{1}{1+\epsilon}} \frac{10tv\Lambda_j^{1+\epsilon} + 2(1+\epsilon)\log\left(\frac{2}{\gamma_j}\right) + 2(1+\epsilon)\log\left(\frac{2}{\theta}\right)}{t\Lambda_j} \leq \\
&\leq v^{\frac{1}{1+\epsilon}}(1+\epsilon)\frac{\frac{10v}{1+\epsilon}t\Lambda_j^{1+\epsilon} + 4\log\left(\frac{2}{\gamma_j}\right)}{t\Lambda_j} \leq \\
&\leq 34v^{\frac{1}{1+\epsilon}}(1+\epsilon)\left(\frac{\log\left(\frac{2}{\gamma}\right) + 2\log(\log(e^2 t))}{t}\right)^{\frac{\epsilon}{1+\epsilon}}.
\end{aligned}$$

Noting that $\sum_{j=1}^{\infty} \gamma_j < \gamma$, this yields a tight bound over the width of the $(1-\gamma)$-CS for μ.

Step 5 (Good event characterization) As the width derived in the previous steps is not deterministic, we now characterize a favorable event in which such bound hold simultaneously for all $(1-\gamma)$-CS. For any $(1-\gamma)$-CS of length t, we have that

$$\mathbb{P}\left(w(t,\mu,\gamma) \le 68 v^{\frac{1}{1+\epsilon}} (1+\epsilon) \left(\frac{\log\left(\frac{2}{\gamma}\right) + 2\log(\log(e^2 t))}{t} \right)^{\frac{\epsilon}{1+\epsilon}} \right) \ge 1 - \frac{\gamma}{4}.$$

Thus, considering a stream of T samples, the probability of this to be violated for at least one interval of the $(1-\gamma)$-CS is bounded as

$$\begin{aligned} 1 - \mathbb{P}(\mathcal{W}_T) &= \mathbb{P}\left(\exists t \le T : w(t,\mu,\gamma) > 68 v^{\frac{1}{1+\epsilon}} (1+\epsilon) \left(\frac{\log\left(\frac{2}{\gamma}\right) + 2\log(\log(e^2 t))}{t} \right)^{\frac{\epsilon}{1+\epsilon}} \right) \\ &\le \sum_{t=1}^{T} \mathbb{P}\left(w(t,\mu,\gamma) > 68 v^{\frac{1}{1+\epsilon}} (1+\epsilon) \left(\frac{\log\left(\frac{2}{\gamma}\right) + 2\log(\log(e^2 t))}{t} \right)^{\frac{\epsilon}{1+\epsilon}} \right) \\ &\le T\frac{\gamma}{4}. \end{aligned}$$

The event $\mathcal{W}_T$, defined above, represents a good event in which the $(1-\gamma)$-CS starting from $t=1$ have the widths of the single CIs deterministically bounded up until horizon T. Now, we note that if we have t different $(1-\gamma)$-CS of lengths $1,\dots,t$, we define $\mathcal{W}_{1:t} := \bigcap_{i=1}^{t} \mathcal{W}_i$. This event describe the scenario in which all $(1-\gamma)$-CS starting sequentially before t have the widths of all of their CIs bounded. Using another union bound argument, we can see that $\mathbb{P}(\mathcal{W}_{1:t}) \ge 1 - \frac{t(t+1)\gamma}{8}$. Finally, note that $\mathcal{W}_{a:b} \subset \mathcal{W}_{a':b'}$, for every $a' > a$ and $b' < b$. Thus $\mathbb{P}(\mathcal{W}_{a:b}) \ge 1 - \frac{T(T+1)\gamma}{8}$ for every $a,b \in [T]$. Characterizing this event is necessary since `Catoni-FCS-detector` requires that CS widths *well behave*, *i.e.*, they possess a deterministic upper bound.

We also introduce the event $\mathcal{E}_t^T = \left\{ \forall i \in \{t,\dots,T\}, \forall t' \in \{i,\dots,T\} : \mu \in CI_{t'}^{(i)} \right\}$, that represents the scenario in which every $(1-\gamma)$-CS starting from a timestamp greater or equal than t never miscovers the true mean up to time T. By the definition of $(1-\gamma)$-CS, we have $\mathbb{P}(\mathcal{E}_t^T) \ge 1 - (T-t)\gamma$. From now on, we continue by setting $\gamma = \frac{2}{T^3}$.

Step 6 (Verifying condition (17)**)** For every $t \in [T]$, we use the previously defined values for $\{\lambda_i\}_{i=1}^{t}$, γ and θ, and solve inequality (17). We obtain that it is satisfied for every $t \ge 68 \log\left(T^{\frac{1+\epsilon}{\epsilon}}\right) = n_{min}$, which is always true under the theorem's assumptions.

Step 7 (Bounding the detection delay) We are now ready to bound the detection delay of `Catoni-FCS-detector` after a change of magnitude δ happened after t_c samples. Note that we assume t_c to be large enough to satisfy Equation (17). To do so, we leverage the width of the $(1-\gamma)$-CS that has just been derived. Suppose, without loss of generality, that a change point is detected after at most T overall samples. Thus, we work under the events $\mathcal{W}_{1:T}$, $\mathcal{E}_1^{t_c}$, and $\mathcal{E}_{t_c}^T$, defined in Step 5, which hold simultaneously with probability at least $1 - \frac{14}{T}$, and guarantee that the CS widths are always properly bounded and the pre-change mean μ_0 and the post-change mean μ_1 are never miscovered. By assumption, t_c is large enough to ensure

$$w(t_c,\mu_0,\gamma) \le 68 v^{\frac{1}{1+\epsilon}} (1+\epsilon) \left(\frac{3\log(T) + 2\log(\log(e^2 t_c))}{t_c} \right)^{\frac{\epsilon}{1+\epsilon}} \le \frac{\delta}{2}$$

and we have to find an n s.t.:

$$w(n,\mu_1,\gamma) \le 68 v^{\frac{1}{1+\epsilon}} (1+\epsilon) \left(\frac{3\log(T) + 2\log(\log(e^2 n))}{n} \right)^{\frac{\epsilon}{1+\epsilon}} \le \frac{\delta}{2}.$$

We first bound $2\log(\log(e^2 n)) \le 3\log(\log(n))$, that holds under the trivial requirements that $\log(n) \ge 2$ and $T \ge 2$. Moreover, we define $\widetilde{c} := 136(3)^{\frac{\epsilon}{1+\epsilon}} (v)^{\frac{1}{1+\epsilon}}$ and $\widetilde{\delta} = \frac{\delta}{\widetilde{c}}$. Thus, we can find an upper bound on the expected

detection delay n_0 by solving the following:

$$n_0 = \min_{n\geq 1}\left\{\left(\frac{\log(T)+\log(\log(n))}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \leq \widetilde{\delta}\right\}.$$

If $\widetilde{\delta}\geq 1$, then $n_0 \leq \log(T)$. Else, for $\widetilde{\delta}\leq 1$, we define

$$n_1 = \min_{n\geq 1}\left\{\left(\frac{\log(\log(n))}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \leq \frac{\widetilde{\delta}}{2}\right\} \quad \text{and} \quad n_2 = \min_{n\geq 1}\left\{\left(\frac{\log(T)}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \leq \frac{\widetilde{\delta}}{2}\right\},$$

and note that $n_0 \leq n_1 + n_2$. We can thus upper bound them separately. It is trivial to observe that

$$n_2 = 2^{\frac{1+\epsilon}{\epsilon}}\frac{\log(T)}{\widetilde{\delta}^{\frac{1+\epsilon}{\epsilon}}} = (2\widetilde{c})^{\frac{1+\epsilon}{\epsilon}}\frac{\log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}}.$$

Upper bounding n_1 requires additional effort. We start by identifying a value n_3 which satisfies

$$\left(\frac{\log(\log(n_3))}{n_3}\right)^{\frac{\epsilon}{1+\epsilon}} \leq \frac{\widetilde{\delta}}{2}.$$

Let $n_3 = \left(\frac{4}{\widetilde{\delta}^2}\right)^{\frac{1+\epsilon}{\epsilon}}$, and $\widetilde{y} := \left(\frac{\widetilde{\delta}}{2}\right)^{\frac{1+\epsilon}{\epsilon}} \leq 1$, then

$$\begin{aligned}
\left(\frac{2}{\widetilde{\delta}}\left(\frac{\log(\log(n_3))}{n_3}\right)^{\frac{\epsilon}{1+\epsilon}}\right)^{\frac{1+\epsilon}{\epsilon}} &= \left(\frac{2}{\widetilde{\delta}}\right)^{\frac{1+\epsilon}{\epsilon}}\left(\frac{\log(\log(n_3))}{n_3}\right) \\
&= \left(\frac{\widetilde{\delta}}{2}\right)^{\frac{1+\epsilon}{\epsilon}}\log\left(\log\left(\left(\frac{4}{\widetilde{\delta}^2}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\right) \\
&= \left(\frac{\widetilde{\delta}}{2}\right)^{\frac{1+\epsilon}{\epsilon}}\log\left(\log\left(\left(\frac{2}{\widetilde{\delta}}\right)^{2\frac{1+\epsilon}{\epsilon}}\right)\right) \\
&= \widetilde{y}\log\left(\log\left(\frac{1}{\widetilde{y}^2}\right)\right) \leq 0.27 < 1.
\end{aligned}$$

Since n_3 is an upper bound on the expected detection delay, we have

$$\begin{aligned}
\log(\log(n_1)) &\leq \log(\log(n_3)) \\
&= \log\left(\log\left(\left(\frac{4}{\widetilde{\delta}^2}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\right) \\
&= \log\left(2\frac{1+\epsilon}{\epsilon}\log\left(\frac{2}{\widetilde{\delta}}\right)\right) \\
&= \log\left(2\frac{1+\epsilon}{\epsilon}\right) + \log\left(\log\left(\frac{2\widetilde{c}}{\delta}\right)\right) \\
&\leq \log\left(2\frac{1+\epsilon}{\epsilon}\right) + \log(\log(2\widetilde{c})) + \log\left(\log\left(\frac{1}{\delta}\right)\right) \\
&= \log\left(2\log(2\widetilde{c})\frac{1+\epsilon}{\epsilon}\right) + \log\left(\log\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

As a consequence, we can rewrite:

$$\left(\frac{\log(\log(n_1))}{n_1}\right)^{\frac{\epsilon}{1+\epsilon}} \leq \left(\frac{\log\left(2\log(2\widetilde{c})\frac{1+\epsilon}{\epsilon}\right) + \log\left(\log\left(\frac{1}{\delta}\right)\right)}{n_1}\right)^{\frac{\epsilon}{1+\epsilon}},$$

which immediately implies that

$$n_1 \leq \frac{(2\widetilde{c})^{\frac{1+\epsilon}{\epsilon}} \log\left(2\log(2\widetilde{c})\frac{1+\epsilon}{\epsilon}\right) + (2\widetilde{c})^{\frac{1+\epsilon}{\epsilon}} \log\left(\log\left(\frac{1}{\delta}\right)\right)}{\delta^{\frac{1+\epsilon}{\epsilon}}}.$$

Under the events defined above and that hold with probability at least $1 - \frac{14}{T}$, the detection delay is bounded as

$$\begin{aligned}(\tau - t_c)^+ &\leq (2\widetilde{c})^{\frac{1+\epsilon}{\epsilon}} \frac{\log\left(2\log(2\widetilde{c})\frac{1+\epsilon}{\epsilon}\right) + \log\left(\log\left(\frac{1}{\delta}\right)\right) + \log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}} \\ &\leq 6(472)^{\frac{1+\epsilon}{\epsilon}} v^{\frac{1}{\epsilon}} \frac{\log\left(\log\left(\frac{1}{\delta}\right)\right) + \log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}}.\end{aligned}$$

Step 8 (Bounding the probability of false alarm) The bound on the probability of false alarm is a trivial consequence of the definition of event $\mathcal{E}_1^{t_c}$. Under this event, it is impossible by construction for the detector to raise a false alarm, as all the CS always intersect at least on μ_0. Thus, the probability of false alarm is bounded by $\mathbb{P}\left((\mathcal{E}_1^{t_c})^C\right) \leq \frac{1}{T}$. □

Lemma A.2. *Let* $\mathcal{G}_T := \left\{\forall j \in [\Upsilon] : \tau_j \in \left\{t_c^{(j)}, \dots, t_c^{(j)} + \left\lceil L_j \frac{K}{\eta} \right\rceil\right\} \text{ and } t_c^{(\Upsilon+1)} > T\right\}$ *be the event in which* `R-CPD-UCB` *restarts exactly* Υ *times without false alarms and excessive delays. Then, we have* $\mathbb{P}\left(\mathcal{G}_T^C\right) \leq \frac{15K\Upsilon}{T}$.

Proof. Note that, by construction of the algorithm, each action is sampled at least L_j times after $\left\lceil L_j \frac{K}{\eta} \right\rceil$ timesteps have passed since the last detection point. Thanks to Assumption 4.7, the length of every epoch is at least $2\left\lceil L_j \frac{K}{\eta} \right\rceil$.

Let $\mathcal{L}^{(j)} := \left\{\forall m \leq j : \tau_m \in \left\{t_c^{(m)}, \dots, t_c^{(m)} + \left\lceil L_m \frac{K}{\eta} \right\rceil\right\}\right\}$ be the event in which all detections happened without false alarms and excessive delays up to the j-th epoch. Then, by a union bound and by Proposition 4.2, we have:

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}_T^C\right) &\leq \sum_{j=1}^{\Upsilon+1} \mathbb{P}\left(\tau_j \leq t_c^{(j)} \mid \mathcal{L}^{(j-1)}\right) + \sum_{j=1}^{\Upsilon} \mathbb{P}\left(\tau_j \geq t_c^{(j)} + \left\lceil L_j \frac{K}{\eta} \right\rceil \mid \mathcal{L}^{(j-1)}\right) \\ &\leq \frac{K(\Upsilon+1)}{T} + \frac{14K\Upsilon}{T} \leq \frac{15K\Upsilon}{T}.\end{aligned}$$

□

Theorem 4.8 (Regret Upper Bound of `R-CPD-UCB`)**.** *Under Assumption 4.7,* `R-CPD-UCB` *suffers an expected cumulative regret bounded as:*

$$\begin{aligned}\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \mathcal{O}\Bigg(&\underbrace{v^{\frac{1}{\epsilon}} \sum_{j=1}^{\Upsilon} \frac{\log(T)}{\left(\widetilde{\delta}_{min}^{(j)}\right)^{\frac{1+\epsilon}{\epsilon}}} \left\lceil \frac{K}{\eta} \right\rceil \Delta_{max}^{(j)}}_{\text{(A) Detection Delay Contribution}} + \\ &+ \underbrace{\sum_{j=1}^{\Upsilon} \mathbb{E}[R^{\pi_s}(|E_j|)]}_{\text{(B) Stationary Policy Regret}} + \underbrace{\eta \sum_{j=1}^{\Upsilon} |E_j| \Delta_{max}^{(j)}}_{\text{(C) Uniform Exploration}} \Bigg).\end{aligned} \tag{6}$$

Proof. Let $\mathcal{G}_T := \left\{\forall j \in [\Upsilon] : \tau_j \in \left\{t_c^{(j)}, \dots, t_c^{(j)} + \left\lceil L_j \frac{K}{\eta} \right\rceil\right\} \text{ and } t_c^{(\Upsilon+1)} > T\right\}$ be the event in which `R-CPD-UCB` restarts exactly Υ times without false alarms and excessive delays.

We start by decomposing the regret in the following way:

$$\begin{aligned}\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] &\leq \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) \mid \mathcal{G}_T] + \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) \mid \mathcal{G}_T^C]\mathbb{P}\left(\mathcal{G}_T^C\right) \\ &\leq \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) \mid \mathcal{G}_T] + 15K\Upsilon,\end{aligned}$$

where the second inequality follows from Lemma A.2. We can now focus on bounding the first addendum. We decompose it as follows:

$$\begin{aligned}\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) \mid \mathcal{G}_T] &= \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) - R^{\pi^{\texttt{R-CPD-UCB}}}(t_c^{(1)}) \mid \mathcal{G}_T] + \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(t_c^{(1)}) \mid \mathcal{G}_T] \\ &\leq \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) - R^{\pi^{\texttt{R-CPD-UCB}}}(t_c^{(1)}) \mid \mathcal{G}_T] + \eta t_c^{(1)} \Delta_{max}^{(1)} + \mathbb{E}[R^{\pi_s}(t_c^{(1)})],\end{aligned}$$

where the inequality follows by upper bounding the contribution to the regret given by the forced exploration in the first $t_c^{(1)}$ rounds, the remaining term is the expected regret accrued by the policy π_s up to $t_c^{(1)}$. We prosecute by bounding the first addendum as follows:

$$\begin{aligned}\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) - R^{\pi^{\texttt{R-CPD-UCB}}}(t_c^{(1)}) \mid \mathcal{G}_T] &= \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) - R^{\pi^{\texttt{R-CPD-UCB}}}(\tau_1) \mid \mathcal{G}_T] + \\ &\quad + \mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(\tau_1) - R^{\pi^{\texttt{R-CPD-UCB}}}(t_c^{(1)}) \mid \mathcal{G}_T] \\ &\leq \mathbb{E}_2[R^{\pi^{\texttt{R-CPD-UCB}}}(T - \tau_1) \mid \mathcal{G}_T] + \mathbb{E}[(\tau_1 - t_c^{(1)}) \mid \mathcal{G}_T] \Delta_{max}^{(1)},\end{aligned}$$

where $\mathbb{E}_2$ is the expectation according to an environment starting from the second segment. Putting all together, we can write

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) \mid \mathcal{G}_T] \leq \mathbb{E}_2[R^{\pi^{\texttt{R-CPD-UCB}}}(T - \tau_1) \mid \mathcal{G}_T] + \mathbb{E}[(\tau_1 - t_c^{(1)}) \mid \mathcal{G}_T] \Delta_{max}^{(1)} + \eta t_c^{(1)} \Delta_{max}^{(1)} + \mathbb{E}[R^{\pi_s}(t_c^{(1)})],$$

which yields, by a recursive application:

$$\begin{aligned}\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T) \mid \mathcal{G}_T] &\leq \sum_{j=1}^{\Upsilon} \mathbb{E}[(\tau_j - t_c^{(j)}) \mid \mathcal{G}_T] \Delta_{max}^{(1)} + \sum_{j=1}^{\Upsilon} \mathbb{E}[R^{\pi_s}(|E_j|)] + \eta T \Delta_{max}^{(1)} \\ &\leq \sum_{j=1}^{\Upsilon} \left\lceil L_j \frac{K}{\eta} \right\rceil \Delta_{max}^{(1)} + \sum_{j=1}^{\Upsilon} \mathbb{E}[R^{\pi_s}(|E_j|)] + \eta T \Delta_{max}^{(1)},\end{aligned}$$

where the second inequality follows from the definition of $\mathcal{G}_T$. The proof is concluded by substituting L_j with its definition. □

Corollary 4.10. *Let π_s be the* `Robust UCB` *policy with median-of-means estimator (Bubeck et al., 2013, Section 2.2). Under Assumption 4.7,* `R-CPD-UCB` *suffers an expected cumulative regret bounded as:*

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \mathcal{O}\left((A) + \underbrace{\sum_{j=1}^{\Upsilon} \sum_{i:\Delta_i^{(j)}>0} \frac{v^{\frac{1}{\epsilon}} \log(|E_j|)}{\left(\Delta_i^{(j)}\right)^{\frac{1}{\epsilon}}}}_{\substack{(B_1)\ \textit{Robust UCB Regret} \\ \textit{(Instance Dependent)}}} + (C) \right). \tag{9}$$

Moreover, if $\log(|E_j|) \geq \frac{5(\Delta_{max}^{(j)})^{\frac{1+\epsilon}{\epsilon}}}{2v^{\frac{1}{\epsilon}}}$ *for every* $j \in [\Upsilon]$*, we have*

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \widetilde{\mathcal{O}}\left((A) + \underbrace{(K\Upsilon)^{\frac{\epsilon}{1+\epsilon}} (vT)^{\frac{1}{1+\epsilon}}}_{\substack{(B_2)\ \textit{Robust UCB Regret} \\ \textit{(Instance Independent)}}} + (C) \right). \tag{10}$$

Proof. The proof of this theorem trivially follows from plugging the regret bounds of `Robust UCB` with MoM estimator (Theorem 3 and Proposition 1 of Bubeck et al. (2013)). Equation (10) necessitates an additional step using Jensen Inequality:

$$\sum_{j=1}^{\Upsilon} K^{\frac{\epsilon}{1+\epsilon}} (vT)^{\frac{1}{1+\epsilon}} \leq \Upsilon (K)^{\frac{\epsilon}{1+\epsilon}} \left(\frac{vT}{\Upsilon}\right)^{\frac{1}{1+\epsilon}} = (\Upsilon K)^{\frac{\epsilon}{1+\epsilon}} (vT)^{\frac{1}{1+\epsilon}}.$$

□

Corollary 4.11. *Let π_s be the* `Robust UCB` *policy with median-of-means estimator from (Bubeck et al., 2013, Section 2.2). Let $\{\eta_j\}_{j\in\mathbb{N}}$ where $\eta_j = \eta_0\sqrt{jK\log(T)/T}$ for some $\eta_0 > 0$. Under Assumption 4.7,* `R-CPD-UCB` *using η_{j+1} after the j-th detection suffers an expected cumulative regret bounded as:*

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \mathcal{O}\left(\frac{v^{\frac{1}{\epsilon}}\sqrt{K\Upsilon T\log(T)}}{\eta_0 \delta_{min}{}^{\frac{1+\epsilon}{\epsilon}}}\Delta_{max} + \frac{K\Upsilon v^{\frac{1}{\epsilon}}\log(T/\Upsilon)}{\Delta_{min}^{\frac{1}{\epsilon}}}\right). \tag{11}$$

Moreover, if $\log(|E_j|) \geq 3(\Delta_{max}^{(j)})^{\frac{1+\epsilon}{\epsilon}} v^{-\frac{1}{\epsilon}}$ for every $j \in [\Upsilon]$, and $\delta_{min}^{\frac{1+\epsilon}{\epsilon}} \geq v^{\frac{1}{\epsilon(1+\epsilon)}}(\Upsilon K/T)^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\log(T)}$, we have

$$\mathbb{E}[R^{\pi^{\texttt{R-CPD-UCB}}}(T)] \leq \widetilde{\mathcal{O}}\Big((K\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}\Big). \tag{12}$$

Proof. First, note that the proof of Theorem 4.8 can be conducted in the exact same way by substituting η with the sequence $\{\eta_j\}_{j\in[\Upsilon+1]}$. Note that, thanks to event $\mathcal{G}_T$ the algorithm restarts exactly Υ times. Equation (8) is a trivial consequence of the fact that $\eta_{\Upsilon+1} \geq \eta_j$ for every $j \leq \Upsilon$, due to the monotonocity of the sequence. Moreover, we bound $\Delta_{max}^{(j)} \leq \Delta_{max}$.

To prove Equation (**??**), we need an additional step. In particular:

$$\sum_{j=1}^{\Upsilon}\frac{1}{\eta_j} = \frac{1}{\eta_0}\sqrt{\frac{T}{K\log(T)}}\sum_{j=1}^{\Upsilon}\frac{1}{\sqrt{j}} \leq \frac{1}{\eta_0}\sqrt{\frac{\Upsilon T}{K\log(T)}}.$$

Plugging this in (A), and bounding $\widetilde{\delta}_{min}^{(j)} \geq \delta_{min}$ for every $j \in [\Upsilon]$, concludes the proof. □

B Additional Related Works on Non-Stationary MABs

In this appendix, we discuss more in detail the related works on non-stationary MABs.

B.1 Piecewise-Stationary MABs

The most common definition of piecewise-stationary MABs in the literature is the one introduced by Yu and Mannor (2009). In this work, the authors deal with the PS MAB problem as it is defined in this work, and consider both a scenario in which side-observations are available, and an agnostic scenario in which they are not, which corresponds to the one that we study in this work. In the latter scenario, they show that every algorithm must suffer at least $\Omega(\sqrt{T})$ regret. In Garivier and Moulines (2011), the authors analyze two algorithms to tackle the PS MAB problem, namely `Discounted UCB` (introduced in Kocsis and Szepesvári (2006)) and `Sliding Window UCB`. Contrary to ours, these algorithms don't rely on any CPD strategy but rather *passively* adapt to the changes in the environment. In practice, *actively* adapting algorithm, *e.g.*, algorithm based on CPD strategies like ours, hence show better performances. The idea of actively adapt to changes first appeared in Hartland et al. (2007). More recent works, such as Liu et al. (2018) and Cao et al. (2019), paved the way for the analysis of actively adaptive algorithms, which were considered tougher to analyze from a theoretical perspective w.r.t. to their passive counterparts. Recently, with Auer et al. (2019) and Besson et al. (2022), there has been focus on removing the prior knowledge on Υ from the algorithms. In the former, the `AdSwitch` algorithm they propose does not require any additional assumptions, but it is not optimized for tractability or numerical efficiency. Indeed, as shown in Besson et al. (2022), `AdSwitch` enjoys poor empirical performance. In the latter, the authors propose an algorithm that performs well in practice and has tight theoretical guarantees without any need for Υ to be known beforehand, however they rely on an assumption which is nearly the same as ours. All of the aforementioned works don't account for the heavy-tailed setting, as their scope is restricted to rewards with bounded support or subgaussian. The only work that accounts for non-stationarity in heavy-tailed settings is Bhatt et al. (2023), where the authors consider a general framework to allow for more general risk measures (linear being the case considered here), and consider the same setup of piecewise-stationary bandits and heavy-tailed rewards, and establish upper and lower bounds on the regret under special assumptions on the risk measures and distributions. However, there are multiple reasons for why our approach is better suited in the regret-minimization scenario:

- Assumption 1 (Stability): since their paper focuses on achieving strong regret guarantees with heavy-tailed rewards, the stability assumption plays a crucial role in the analysis, where the rate functions on the decay of the empirical and truncated distributions are assumed to be known. While the assumption is not strong in and of itself, the knowledge of these parameters play a crucial role in the change detection and regret minimization procedures, and also appear in the regret bounds. `Robust-CPD-UCB`, on the other hand, requires no knowledge of such functions, relies on novel analysis of Catoni estimators that do not necessitate truncations, and is simpler to run online in practice.
- The CPD routine used in Bhatt et al. (2023) is based on a sliding window method requiring specification of both widow size and threshold. `Robust-CPD-UCB` is based on the newly developed CPD method based on Catoni estimator that runs online only with the same assumption on the distributions.
- The regret minimization algorithm in Bhatt et al. (2023) uses a data-driven truncation of the rewards that depends on the policy, and the knowledge of the decay rate functions to compute the exploration bias for the arm index. Our work, on the other hand, requires no such methods and uses a simple combination of the novel Catoni-CPD and any policy suited for the stationary heavy-tailed regret minimization.

For the most common case of linear risk/ regret in mean, `Robust-CPD-UCB` establishes stronger guarantees with the CPD procedure requiring weaker assumptions. Finally, it is easier to implement owing to not requiring distributional knowledge or thresholds.

B.2 Bounded Variation and Monotonically Non-stationary MABs

Another setting of interest is non-stationary MABs with *bounded variations*. In this setting, the rewards' distributions changes are less restricted, and the focus moves from the number of changes to the total amount of change V_T. In Besbes et al. (2014), the authors propose `Rexp3`, an algorithm that leverages tools from the adversarial MAB problem to deal with non-stationarity in stochastic settings. The regret upper bound that they provide is in the order of $\mathcal{O}(V_T^{\frac{1}{3}} T^{\frac{2}{3}})$. Over the last years, there has been increasing interest in *monotonically non-stationary* MABs, *i.e.*, non-stationary MABs where the mean rewards are only allowed to decrease (rotting bandits, Seznec et al. (2019, 2020)) or to increase (rising bandits, Metelli et al. (2022)), some works focus on both settings Heidari et al. (2016); Genalti et al. (2024b). The monotonicity assumptions substitutes the need for piecewise-stationarity, as it is a strong enough assumption to allow for strong theoretical characterizations. In such settings, regret bounds depend in general on the total variation of the distributions' means and instance-dependent-type of results are common in this literature. Moreover, the additional structure added by this assumption, put the accent on the difference between *restless* bandits (a proper non-stationary setting) and *rested* bandits, where the evolution of rewards depends on learner's actions rather than just time.

C Additional Numerical Evaluations

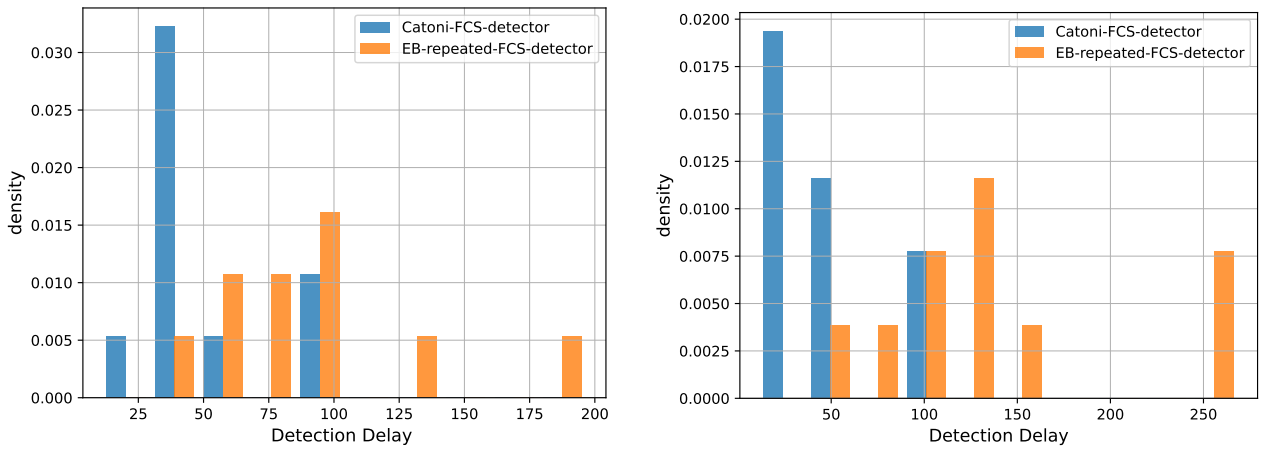
In this appendix, we provide additional details on the experimental evaluation of Section 5 and additional experimental campaigns in synthetic environments.

C.1 Detection Delay Analysis

We evaluate how reactive is `Catoni-FCS-detector` to changes of data-generating distribution, and comparing it `repeated-FCS-detector` with Empirical Bernstein CSs from Shekhar and Ramdas (2023b), Section 2.2, which is suited for distributions with finite variance. We consider two distribution-shift scenarios: Gaussian distributions with $\sigma = 1$ and Laplace distributions with scale equal to 1. The change happens after $t_c = 400$ steps, and the total horizon is $T = 1000$. The magnitude of change is $\delta = 1$. In Figure 4, we report the distribution of the detection delay of both algorithm over 20 trials. We can see how, in general, `Catoni-FCS-detector` has a smaller detection delay w.r.t. `repeated-FCS-detector`. Moreover, no false alarm is raised along the 20 trials.

C.2 Regret Minimization in Highly Non-Stationary Environments

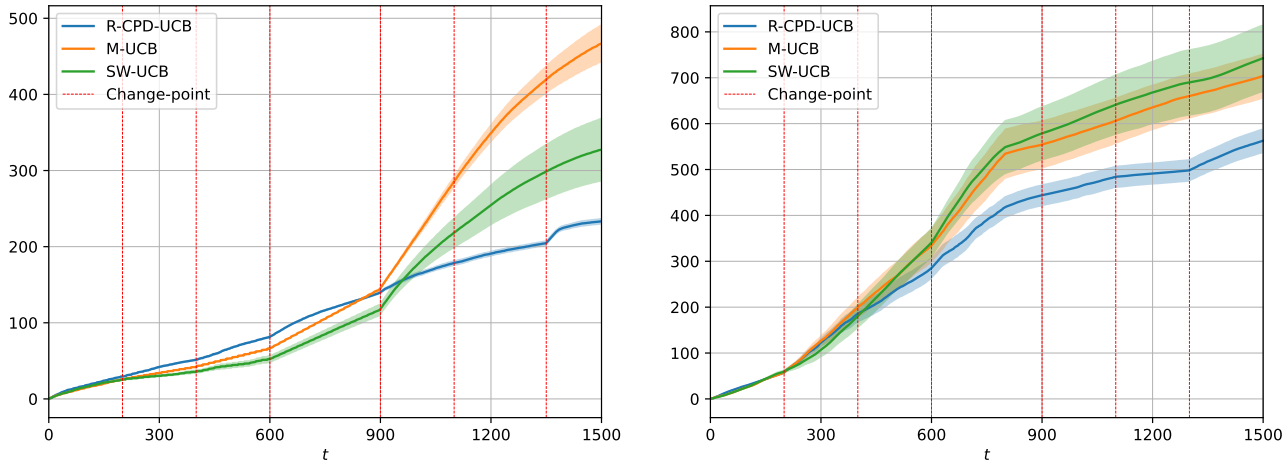
In this section, we evaluate how `R-CPD-UCB` behaves in highly dynamic scenarios where change-points are close.
Setting We confront `R-CPD-UCB` with two of the most popular algorithms from the literature, `Monitored UCB`



(a) Gaussian samples.

(b) Laplace samples.

Figure 4: Distribution delay distribution over 20 trials.



(a) Gaussian rewards.

(b) Pareto rewards.

Figure 5: Cumulative regrets of the considered algorithms. We performed 20 trials for each instance and reported mean $\pm$ std. The 6 change-points are indicated by the vertical lines.

Cao et al. (2019) and `Sliding Window UCB` Garivier and Moulines (2011). We consider two PS MABs: Gaussian rewards with $\sigma = 1$ and Pareto rewards with $\epsilon < \frac{1}{2}$ and $v < 3$. In both MABs, we have $K = 3$, $T = 1500$ and $\Upsilon = 6$. Also, we have $\delta_{min} = 0$, *i.e.*, and some actions may not change their means after a change-point. However, at least one arm has its mean change, and the optimal action changes at least 4 times. Interestingly, these instances violate Assumption 4.7. We use $\sigma = 1$ and the means reported in Table 6a for the Gaussian scenario. The optimal actions change 4 times. For the Pareto scenario, we use $\epsilon = \frac{1}{2}$, $v = 3$, and the means reported in Table 6b. The optimal actions change 4 times. In Figure 7 we report the means of every action in every epoch.

	E_1	E_2	E_3	E_4	E_5	E_6	E_7
μ_1	1.2	1.5	1.5	2	1.8	1.2	1.2
μ_2	1	1.8	2.4	1.8	1	1.8	1
μ_3	0.5	0.5	0.5	0.5	1	0.5	$1, 7$

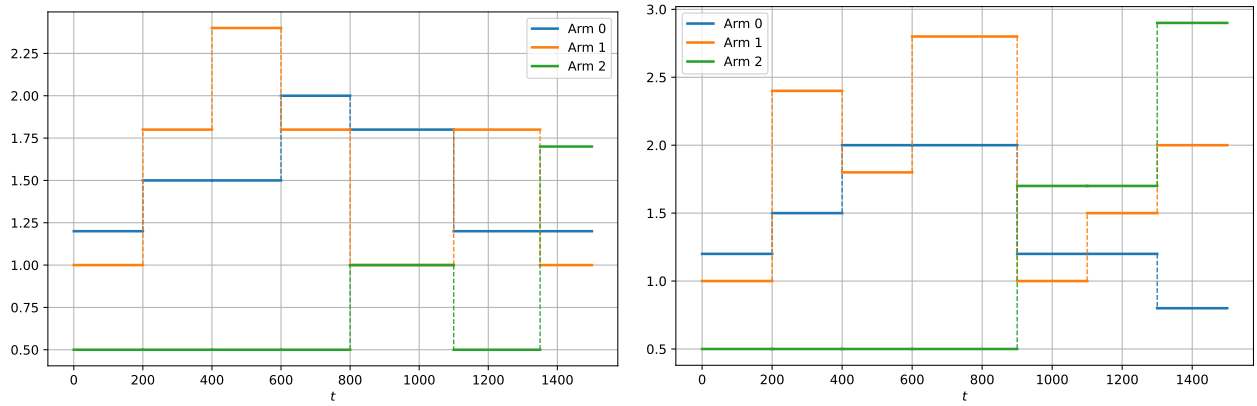
(a) Gaussian instance.

	E_1	E_2	E_3	E_4	E_5	E_6	E_7
μ_1	1.2	1.5	2	2	1.2	1.2	0.8
μ_2	1	2.4	1.8	2.8	1	1.5	2
μ_3	0.5	0.5	0.5	0.5	1.7	1.7	2.9

(b) Pareto instance.

Figure 6: Mean rewards per epoch. Cells highlighted in yellow contain the optimal reward for the corresponding epoch.

Results In Figure 5, we report the cumulative regrets suffered by the considered algorithms. For each instance and algorithm, we performed 20 trials and reported the average cumulative regrets with their aleatoric uncertainties. `R-CPD-UCB` achieves, in both instances, a smaller cumulative regret than competitors. Moreover, it shows a smaller uncertainty and more stable performances across the trials, especially when rewards have infinite variance (Figure 5b). Interestingly, `R-CPD-UCB` can outperform both `Monitored UCB` and `Sliding Window UCB` even when the rewards are Gaussian. This is probably because the change-points are frequent and very close. Robust mean estimation using median-of-means stabilizes the algorithm's behavior in data-scarce regimes. Finally, we remark that Assumption 4.7 is violated by these two instances; however, `R-CPD-UCB` performs well (and so is `Monitored UCB`, which relies on a similar hypothesis). This phenomenon was already observed in Cao et al. (2019), and shows how Assumption 4.7 is, in practice, is not very limiting.



(a) Gaussian rewards.

(b) Pareto rewards.

Figure 7: Average rewards per epoch.

C.3 Sensibility to δ

In this section, we study the sensibility of `R-CPD-UCB` to different magnitudes of changes.

Setting We consider four HTPS MABs with Pareto rewards ($\epsilon < 1$, $v = 1$), $K = 3$, $T = 500$, and $\Upsilon = 1$. The starting means are $\mu_1 = 1$, $\mu_2 = 0.8$ and $\mu_2 = 0.5$, and a change occurs at $t_c = 200$. We let $\delta_1 = 1$, $\delta_3 = 0$ (thus, $\delta_{min} = 0$) and $\delta_2 \in \{1, 2, 5, 10\}$, respectively. In the first PS MAB, the first action remains optimal from the start to the end of the trial; in the others, the second action becomes optimal after the change. In Figure 8, we report the means of every action in every epoch for the four HTPS MABs.

Results In Figure 9, we report the cumulative regrets suffered by `R-CPD-UCB` in the four HTPS MABs. For each instance, we performed 20 trials and reported the average cumulative regret (on the right), together with its

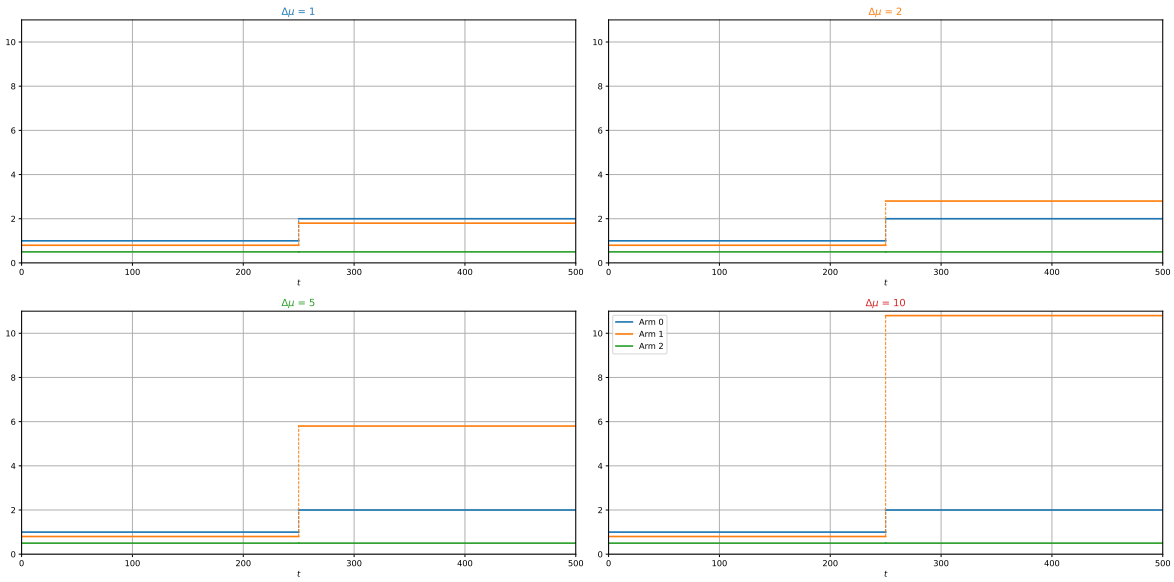


Figure 8: HTPS MABs with different magnitudes of change. For all instances: $\Upsilon = 1$, $K = 3$, $\delta \in \{1, 2, 5, 10\}$. Pareto noise with $\epsilon < 1$ and $v = 3$.

	μ_1	μ_2	μ_3
Instance 1	1	0.5	0.1
Instance 2	1	0.8	0.7
Instance 3	1	0.9	0.1
Instance 4	1	0.5	0.5

Table 1: Mean rewards per epoch in four stationary HT MABs. For all instances: $\Upsilon = 0$, $K = 3$. Pareto noise with $\epsilon < 1$ and $v = 3$.

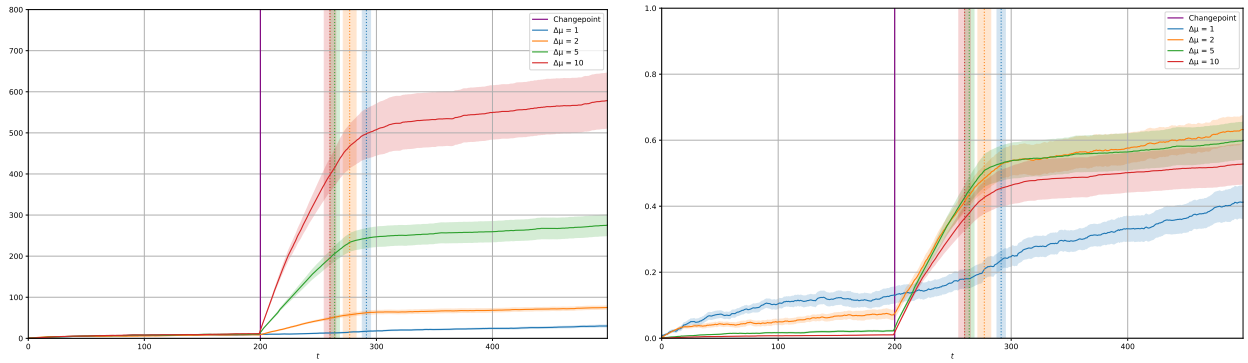
standard deviation. Moreover, the dashed vertical lines indicate the average detection time and their standard deviations. As the four instances differ in terms of magnitude (*e.g.*, in the first instance, the maximum mean is 2, and in the fourth is 10.8), we also reported the rescaled cumulative regrets (on the left). `R-CPD-UCB` can detect the change with a reasonable delay, and all cumulative regrets show sublinear growth. As δ grows, the cumulative regret is larger, but the detection delay decreases. Intuitively, a larger change yields a larger regret but is also easier to detect. From the rescaled cumulative regrets, we can observe how a large δ w.r.t. to the mean does not deteriorate the performance of `R-CPD-UCB`.

C.4 Stationary Environments

In this section, we study the empirical behavior of `R-CPD-UCB` in stationary HT MABs.

Setting We consider four HT MABs with Pareto rewards ($\epsilon < 1$, $v = 1$), $K = 3$, $T = 300$, and $\Upsilon = 0$. We compare `R-CPD-UCB` with two gold standards from the literature: `Robust UCB` Bubeck et al. (2013) and `MR-APE` Lee and Lim (2022).

Results We remark that, in a stationary environment, the behavior of `R-CPD-UCB` diverges from the one of `Robust UCB` in only two cases: (i) when there is a false detection (happens with probability smaller than T^{-1}) and (ii) `R-CPD-UCB` performs a forced exploration (once every $\mathcal{O}(T^{-\frac{1}{2}})$ rounds). In both cases, the contribution to the regret is small compared to the dominant term, adding a constant factor at most. In Figure 10, we report the cumulative regrets suffered by the algorithms in the four HT MABs. For each instance, we performed 20 trials and reported the average cumulative regret, together with its standard deviation. `R-CPD-UCB` raises only one false alarm in one trial of the fourth instance, and the average cumulative regret is thus slightly larger than the one `Robust UCB`, which is suited for the stationary setting. `R-CPD-UCB` suffers cumulative regrets comparable to the ones of algorithms suited for the stationary case. All cumulative regrets show sublinear growth.



(a) Cumulative Regret.

(b) Rescaled Cumulative Regret.

Figure 9: Cumulative regrets of `R-CPD-UCB` in the four HTPS MABs represented in Figure 6. We performed 20 trials for each instance and reported mean ± std. The purple vertical line indicates the change-point. The dashed vertical lines indicated the average detection time in the corresponding instance (± std). On the left, we report the cumulative regrets of `Robust-CPD-UCB`. On the right, we have the same quantity rescaled by the maximum mean reward.

D Computational Complexity of `Robust-CPD-UCB`

In this section, we characterize the computational complexity of `R-CPD-UCB` and provide a simple modification to the algorithm that allows for quicker computation while keeping similar theoretical guarantees. For the sake of traceability, we assume that all means belong to the set $[-M, M]$. We start by upper bounding the computational complexity of `R-CPD-UCB`.

Proposition D.1. *Let* $\widetilde{M} = \mathcal{O}\left(M + v^{\frac{1}{1+\epsilon}} \log(T)^{\frac{\epsilon}{1+\epsilon}}\right)$. *Let* ξ *be the machine tolerance. Then,* `R-CPD-UCB` *takes at most* $\mathcal{O}\left(KT^4 \log_2\left(\frac{\widetilde{M}}{\xi}\right)\right)$ *steps with probability at least* $1 - \frac{1}{T}$.

Proof. Using the bisection method with a tolerance of ξ, and searching in the interval $[-\widetilde{M}, \widetilde{M}]$, we can solve the two root-finding problems implied by Equation (5) in at most $\mathcal{O}\left(T \log_2\left(\frac{\widetilde{M}}{\xi}\right)\right)$. Note that, by Theorem 3.2 from Bhatt et al. (2022), the solution lies in the search interval with probability at least $\Omega\left(1 - \frac{1}{T}\right)$. Then, we observe that `R-CPD-UCB`, at each round $t \in [T]$, for every action $i \in [K]$, runs a step of the `Catoni-FCS-detector` which computes t CS of lengths $t, t-1, \ldots, 2, 1$. Computing a CS of length $t' \leq T$ requires to compute t' CIs, which requires $\mathcal{O}\left(t' \log_2\left(\frac{\widetilde{M}}{\xi}\right)\right)$ steps at most for each of them. Note that a solution always exists as the number of samples is always greater than n_{min} when the CPD routine is executed. The result follows by upper bounding $t, t' \leq T$. □

Proposition D.1 states that the computational complexity of `R-CPD-UCB` is polynomial in T.

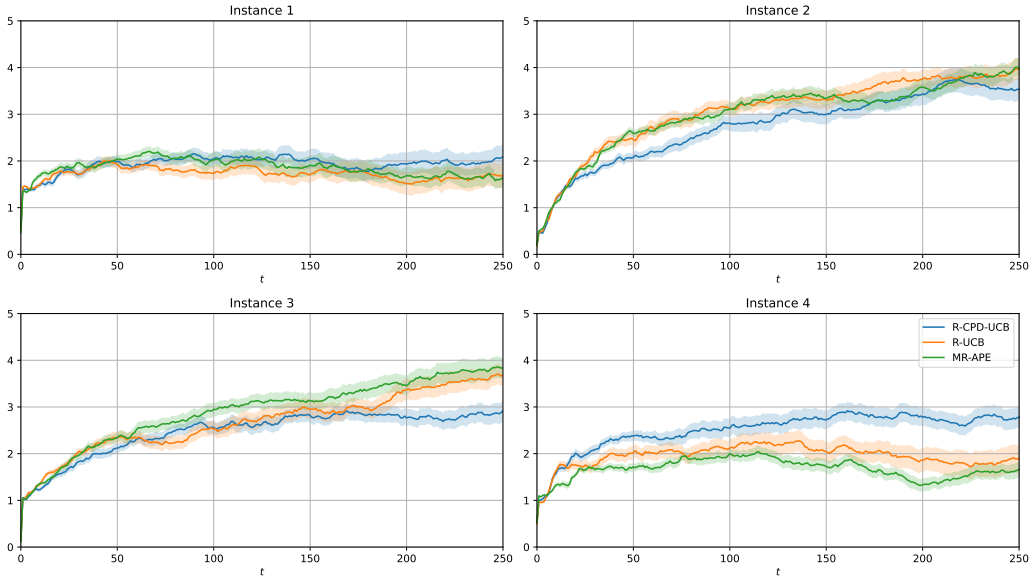


Figure 10: Cumulative regrets of `R-CPD-UCB`, `Robust UCB` (Bubeck et al., 2013), and `MR-APE` (Lee and Lim, 2022) on the four HT MABs defined in Table 3. We performed 20 trials for each instance and reported mean $\pm$ std.