POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA

Multi-Armed Bandits in Dynamic Environments and Heavy-Tailed Rewards

Doctoral Dissertation of:
Gianmarco Genalti

Advisor: Prof. Nicola Gatti
Co-advisor: Prof. Nicolò Cesa-Bianchi
Tutor: Prof. Stefano Ceri
Year 2026 - XXXVIII Cycle

Abstract

The Multi-Armed Bandit (MAB) problem is a theoretical framework that models a broad range of sequential decision-making problems. Over the last two decades, the MAB framework has experienced a rapid growth in interest from both the theoretical research community and practitioners: on one hand, there are several technical challenges posed by proving theoretical guarantees on algorithms; on the other hand, MAB algorithms have been used in relevant real-world applications such as dynamic pricing and digital advertising.

In this work, we examine some of the most significant streams of research in MABs, primarily from a theoretical perspective. In particular, we aim to relax some of the core assumptions of the MAB framework, making it more suited for real-world scenarios, and provide algorithms with provable theoretical guarantees.

We mainly focus on the *regret minimization* problem, where the decision-maker observes a realization of the reward of an action after having chosen it, and aims at maximizing the overall total reward at the end. We aim at bridging the MAB problem with Markov Decision Process (MDP) problem, and to provide MAB-style algorithm with provable theoretical guarantees in the latter. Such settings do not allow for trivial characterizations of the optimal policy, allowing past actions to affect the present.

The largest literature on regret minimization in MABs deals with stochastic realizations that come from *fixed* and *well-behaved* (*e.g.* bounded support or *sub-Gaussian*) probability distributions. In this thesis, we address the *heavy-tailed* bandit problem, where assumptions on the reward-generating distributions are reduced to the bare minimum and the variance may be infinite.

Keywords: multi-armed bandits, regret minimization, heavy-tailed distributions

Sommario

Il problema del *Multi-Armed Bandit* (MAB) rappresenta un quadro teorico che modella un'ampia varietà di problemi di decisione sequenziale. Nel corso degli ultimi due decenni, il modello MAB ha suscitato un crescente interesse sia nella comunità di ricerca teorica sia tra i professionisti: da un lato, vi sono numerose sfide tecniche legate alla dimostrazione di garanzie teoriche sugli algoritmi; dall'altro, gli algoritmi MAB sono stati applicati con successo in importanti contesti reali, come la definizione dinamica dei prezzi e la pubblicità digitale.

In questo lavoro, esaminiamo alcuni dei filoni di ricerca più rilevanti nell'ambito dei MAB, principalmente da una prospettiva teorica. In particolare, ci proponiamo di rilassare alcune delle assunzioni fondamentali del modello, rendendolo più adeguato a scenari reali, e di fornire algoritmi dotati di garanzie teoriche dimostrabili.

Ci concentriamo soprattutto sul problema della minimizzazione del *regret*, in cui il decisore osserva la realizzazione della ricompensa di un'azione dopo averla scelta e mira a massimizzare il totale delle ricompense ottenute al termine dell'orizzonte temporale. Il nostro obiettivo è creare un collegamento tra il problema MAB e quello dei Markov Decision Process (MDP), fornendo algoritmi in stile MAB che mantengano garanzie teoriche anche in quest'ultimo contesto. Tali scenari non consentono, in generale, delle caratterizzazioni banali della politica ottimale, poiché le azioni passate influenzano quelle future.

La maggior parte della letteratura sulla minimizzazione del regret nei MAB considera realizzazioni stocastiche che provengono da distribuzioni di probabilità fisse e molto regolari (ad esempio distribuzioni con supporto limitato o sub-Gaussiane). In questa tesi, affrontiamo il problema degli *heavy-tailed bandits*, in cui le assunzioni sulle distribuzioni che generano le ricompense sono ridotte al minimo indispensabile e la varianza di quest'ultime può essere infinita.

Contents

1 | Introduction

Artificial Intelligence (AI) is now ubiquitous across nearly every aspect of daily life. Many of the digital systems we interact with on a daily basis are capable of making autonomous decisions. Notable examples include e-mail spam filters, recommender systems in streaming platforms, and dynamic pricing in e-commerce platforms. This continuous interaction between humans and autonomous agents has spurred rapid advances in AI capabilities and in the availability of data. At its core, the theory behind AI, particularly for decision-making in uncertain environments, draws from statistics, optimization, and theoretical computer science. Over the past years, the gap and interplay between theory and application in AI have become increasingly evident: practical breakthroughs have inspired rich theoretical research, while sound theoretical understanding remains essential for ensuring the safe and effective deployment of these tools. In the specific paradigm of autonomous agents, where an agent has to make multiple decisions and account for a variety of scenarios, theoretical guarantees are now more important than ever.

1.1. From Batch Learning to Online Learning

In modern learning theory, a major distinction is between *batch learning* and *online learning*. In batch learning, an agent is provided with a static set of data, it learns from them, and is finally asked to provide an output. An output can be, for example, a forecast on the future behavior of the data or to classify a fresh portion of data that is provided afterward. The objective function is only dependent on the quality of the output on the new data. In online learning, an agent is asked to actively collect its own data, learn from it, and provide an optimal behavioral policy for data collection. This time, the objective function of the agent also includes the quality of its behavior during the data collection.

In batch learning, an agent is immutable. Once the first batch of data is provided to the agent, it is never updated afterward. Its theoretical guarantees characterize the error in generalizing to new data what it has learn from the provided data. An agent that successfully captures the underlying data-generating process will, under the assumption of a stable environment, generalize well to new data.

In online learning, the agent can potentially start with no data at all. It is required to collect its own data, update its policy dynamically, and potentially continue forever. In general, its goal may be complex and not limited to collecting the highest quality dataset. For example, an agent may be given a reward every time it does a certain action, not necessarily the one providing the most informative data. This tension between data collection and performance optimization makes the study of online learning both challenging and theoretically rich.

1.2. Thesis Contributions

This thesis advances the theoretical understanding of online learning through the study of Multi-Armed Bandit (MAB) problems and their extensions. These simple, yet powerful, frameworks allow us to provide strong characterizations of the performance of an online learning agent, and to make the tension between *exploration* (collecting informative data points) and *exploitation* (optimizing a given objective function) explicit and manageable.

In what follows, we outline the structure and the contributions of this thesis.

- In Chapter 2, we provide a technical background on Multi-Armed Bandit problems (MABs, Section 2.1) and Markov Decision Processes (MDPs, Section 2.2). We introduce the core assumptions and the relevant literature.
- In Chapter 3, we study some problems that lie between MABs and MDPs.
  - In Section 3.1, we introduce the *Autoregressive Bandit problem*, a particular variant of the MAB setting where rewards are governed by an autoregressive process. We present an algorithm and provide theoretical guarantees on its performance. We validate our approach numerically on synthetic data.
  - In Section 3.2, we introduce the *Graph-Triggered Bandit* problem, a novel framework that bridges *rested* and *restless* bandits, two well-known approaches to model the non-stationarity of the rewards in a bandit problem. We provide a study on the setting complexity from both a statistical and computational perspective. We provide algorithms to deal with both deterministic and stochastic environments, and characterize their performance from a theoretical perspective.
  - In Section 3.3, we focus on restless rising bandits, a widely studied setting over the last years. We address the core open problem of characterizing what is the theoretically optimal performance for an algorithms, and we provide an algorithm that improves over the existing approaches. We numerically validate our approach on synthetic data.

- In Chapter 4, we relax one of the most common assumptions in the MAB literature, and allow for reward distributions that are *heavy-tailed*.
  - In Section 4.1, we address the recent open problem of doing regret minimization in heavy-tailed MABs while being agnostic w.r.t. the parameters of the reward distributions. We are the first to provide impossibility results on such tasks and to provide an algorithm that is optimal under a mild assumption.
  - In Section 4.2, we combine the heavy-tailed MAB problem with the piecewise-stationary bandit problem. Motivated by real-world applications such as finance and telecommunications, we provide the first change-point detection routine for heavy-tailed random variables, and we apply it to regret minimization. We validate our solution with synthetic and real-world data.

2 | Foundations of Online Learning

Online Learning is an umbrella term that encompasses multiple theoretical frameworks in which an autonomous agent sequentially interacts with a partially observable environment and has to learn an optimal behavioral policy, learning from its past decisions. Online learning can be positioned within the broader concept of machine learning, and exhibits many conceptual differences from the largely studied problems of *supervised* and *unsupervised learning*. Indeed, in supervised learning, data are provided in batches, and a learner has the goal to provide a guess on the relationship between a specific segment of information (*i.e.*,, a target variable, for tabular datasets) and the rest of the information available. In online learning, data are *not* provided in batches, but the learner is asked to actively collect them. This element brings a crucial shift from supervised to online learning, as the agent has to learn a behavior while simultaneously understanding the data-generation mechanism.

In this chapter, we introduce the *Multi-Armed Bandit* problem (Section 2.1), one of the most fundamental settings in online learning and the main subject of this thesis. Afterward, we introduce the *Markov Decision Process* problem (Section 2.2) and present it as a generalization of the MAB problem, highlighting the connections between the two and why the results contained in this thesis are relevant to this class of problems.

2.1. The Stochastic Multi-Armed Bandit Problem

In this section, we discuss the Multi-Armed Bandit problem (MAB, Lattimore and Szepesvári (2020)), which constitutes the main theoretical framework for almost all the results presented in this thesis.

The MAB framework is used to address sequential decision-making in problems where feedback is uncertain. In particular, it is one of the most flexible and versatile frameworks to encode a situation in which an autonomous agent, or *learner*, is faced with a repeated number of decisions among a finite set of *actions* (or *arms*), and it is only allowed to observe partial and noisy feedback. Altough MABs have been extensively studied under many different assumptions on the reward generating process, in this thesis we focus on *stochastic* environments, or stochastic MABs. In

Algorithm 1: Learner-Environment Interaction Protocol for Stochastic MABs

1 **for** $t \in [T]$ **do**
2 Learner plays an action $I_t \in [k]$ (possibly at random)
3 Environment receives I_t and provides a reward $X_{I_t,t}$ sampled from ν_{I_t}.
4 Learner receives $X_{I_t,t}$ and updates its internal status.
5 **end**

stochastic MABs, each action is associated to a probability distribution from which a reward is randomly sampled every time the learner plays the associated action. Such environments are popular due to their flexibility, closeness to real-world models and quality of the theoretical guarantees that can be provided over the learner's performance.

Setting Formalization

We define an instance of the stochastic MAB problem (also called k-armed bandit) as a set $\boldsymbol{\nu} = (\nu_i)_{i \in [K]}$ of K probability distributions. The interaction protocol between a learner and an instance of stochastic MAB is exemplified in Algorithm 1. A *trial* lasts $T \in \mathbb{N}$ rounds, and at every round $t \in [T]$ the learner provides a decision $I_t \in [k]$, possibly at random (line 2). The environment receives the decision I_t and samples a reward $X_{I_t,t} \sim \nu_{I_t}$ (line 3). Finally, the learner receives (and observes) the reward and $X_{I_t,t}$ and updates its strategy accordingly (line 4).

We now define some quantities that will come in handy later. First, we call $\mu_i := \mathbb{E}[X_{i,t}]$ the expected reward of action $i \in [k]$. As we will show in the section, we are particularly interested in the action $i^* \in \arg\max_{i \in [k]} \mu_i$, the one having the largest expected reward. Also, let $\mu^* := \mu_{i^*}$. We call $\Delta_{i,j} := \mu_i - \mu_j$ the *gap* between actions i and j. A crucial quantity is the *sub-optimality gap* of action i, defined as $\Delta_i = \mu^* - \mu_i$. We call $N_{i,t} = \sum_{l=1}^{t} -1\mathbb{1}_{\{I_t = i\}}$ the (random) number of times action i has been chosen by the learner before time t. It is customary in the literature to assume that ν_i is a sub-Gaussian probability distribution. A (zero-mean) random variable X is σ-subgaussian if it holds $\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leqslant \exp\left(\frac{\sigma^2 \lambda^2}{2}\right)$ for every $\lambda \in \mathbb{R}$. Note that this definition includes all distributions with bounded support. We will make the same assumption in Chapter 3, but not in Chapter 4.

Learning Goal

In this thesis, we investigate the main learning problem for MABs: *regret minimization*. Regret minimization represents the most investigated learning goal in the MAB literature and, possibly, one of the most natural. Informally, the *regret* of an algorithm is the performance gap of the learner w.r.t. the best possible decision maker. As the environment is stochastic, we compare the performance of the two in expectation.

Before formally introducing the notion of regret, we provide a formalization of what *actually* is a learner, that we will make coincide with the notion of *policy*. We define $\mathcal{H}_t = \{(I_l, X_{I_l,l})\}_{l \in [t]}$ as the *history of interactions* at a given round $t \in [T]$. We define a policy $\pi(t)$ as a (possibly randomized) function $\pi(t) : \mathcal{H}_{t-1} \mapsto I_t$ returning the next action given the history up to that round.

For a given instance $\boldsymbol{\nu}$ of a MAB, the performance of a policy π is measured by the means of *expected cumulative reward* throughout T rounds, formally:

$$J_{\boldsymbol{\nu},T}(\pi) := \mathbb{E}\left[\sum_{t \in [T]} \mu_{I_t}\right],$$

where the expectation is taken over the randomness of both the environment and the policy/algorithm. A policy is *optimal* for instance $\boldsymbol{\nu}$ and time horizon T if it maximizes the expected cumulative reward, formally:

$$\pi^*_{\boldsymbol{\nu},T} \in \arg\max_{\pi} J_{\boldsymbol{\nu},T}(\pi).$$

We denote by $J^*_{\boldsymbol{\nu},T} = J_{\boldsymbol{\nu},T}(\pi^*_{\boldsymbol{\nu},T})$ the expected cumulative reward attained by the optimal policy. We can now define the *expected cumulative regret* as:

$$R_{\boldsymbol{\nu},T}(\pi) = J^*_{\boldsymbol{\nu},T} - J_{\boldsymbol{\nu},T}(\pi).$$

Therefore, our learning problem is to find a policy π minimizing the expected policy regret $R_{\boldsymbol{\nu},T}(\pi)$.

Remark 1. *In some settings, the optimal policy may depend on both $\boldsymbol{\nu}$ and the time horizon T. In standard stochastic MABs, this is not the case since, trivally, it holds $\pi^*_{\boldsymbol{\nu},T}(t) = \pi^*_{\boldsymbol{\nu}}(t) = i^*$, for every $t \in [T]$. However, through this thesis, we will explore more complex scenarios where the problem instance is characterized by additional parameters rather than just $\boldsymbol{\nu}$ and T.*

The Exploration-Exploitation Dilemma The MAB problem exhibits a crucial trade-off between information gaining and reward gaining. The partial feedback forces any policy to decouple *exploration* and *exploitation*. To minimize regret, a policy may want to always pull the action that is believed to be better. However, this would preclude receiving enough feedback from the other action and possibly discovering a better one. This tension is called *exploration-exploitation dilemma*, and can be exploited to prove impossibility results in MABs. From an algorithm design perspective, this precludes greedy algorithms from performing well. Cleverly performing exploration becomes crucial for every algorithm that wants to perform well in MABs.

Existing Results

Regret minimization in stochastic MABs has been widely studied since the first half of the last century. It is no surprise that the best (least) possible expected cumulative regret that any policy must suffer has already been well characterized. This type of result is called regret *lower bound*, and it is very important in understanding how truly difficult and instance is.

The quality of a policy is evaluated from the tightest possible *upper bound* that can be provided on its expected cumulative regret. Comparing the regret upper bound of a policy to the setting's regret lower bound tells us if the policy is good or not.

Instance-Dependent vs Worst-Case Bounds We differentiate the nature of existing results in two families: *instance-dependent* bounds and *worst-case* (or *minimax*) bounds. Instance-dependent bounds show us how the regret depends on quantities that are specific to a given instance, and better characterize the learning challenge. In fact, not all instances are equal, and a desirable property for an algorithm is to perform well when an instance is ëasy.̈ Keeping k and T fixed, in stochastic MABs the difficulty of an instance is characterized by the sub-optimality gaps $\{\Delta_i\}_{i\in[k]}$, and by the support/dispersion of the probability distributions generating the rewards, which is σ in the case of sub-gaussian distributions. Instance-dependent bounds capture all of these quantities and, in more complex settings, the other parameters that characterize an instance. Worst-case bounds, on the other hand, are more natural from a game-theoretic and information-theoretical perspective of the problem. Such bounds consider the worst-case scenario for an instance, and usually only depend on T and k.

Regret Lower Bounds We start by reporting an instance-dependent regret lower bound. This result first appeared in the seminal work Lai and Robbins (1985) in an *asymptotic* version, *i.e.*, letting T go to infinity. Later, a finite-time version has been derived in Lattimore and Szepesvári (2020).

Lower bounds are often obtained by considering two different instances on which no *reasonable* policy can simultaneously perform well. A policy is called reasonable if its regret is s.t. $R_{\boldsymbol{\nu},T} \leqslant CT^p$, for $p \in (0,1)$ and $C > 0$, and for every instance $\boldsymbol{\nu}$. We use the notation $\Omega(\cdot)$ to indicate the lower bound of a quantity up to lower order terms.

Theorem 2.1 (Instance-Dependent Regret Lower Bound for Stochastic MABs)**.** *Let $\boldsymbol{\nu}$ be a stochastic MAB with k actions, σ-subgaussian reward distributions, and π be any reasonable*

policy. Then the following holds:

$$R_{\boldsymbol{\nu},T}(\pi) \geqslant \Omega\left(\sum_{i\in[k]:\Delta_i>0} \sigma\left[\frac{(1-p)\ln T + \ln\left(\frac{\Delta_i}{C}\right)}{\Delta_i} + \Delta_i\right]^+\right), \tag{2.1}$$

where the $[\cdot]^+$ *notation indicates the positive part.*

The dependence on T (which is, in general, the most interesting to look at) is logarithmic. Thus, no algorithm can achieve a regret growing slower than a logarithm with T or a constant one. Moreover, Equation (2.1) describes how the sub-optimality gaps characterize the difficulty of an instance. The sub-optimality gaps may be very small, for example, in the order of $\frac{1}{T}$. This would result in the Δ_i term in the denominator blowing up, but at the same time the logarithmic term can be come negative and large in magnitude, shrinking the lower bound up to zero. Intuitively, if the gaps are very small, it is harder to identify the best action, but every time a sub-optimal action is chosen, the paid regret is also very small. If the sub-optimality gaps are very large, it is easier to identify the best action (the denominator shrinks the bound), but every mistake is more costly (and the linear term grows). This trade-off leads to a very natural question: what happens in the worst possible instance? This is addressed by the worst-case lower bound.

Theorem 2.2 (Worst-Case Regret Lower Bound for Stochastic MABs)**.** *Let* $\boldsymbol{\nu}$ *be a stochastic MAB with* k *actions,* σ*-subgaussian reward distributions, and* π *be any policy. Then the following holds:*

$$R_{\boldsymbol{\nu},T}(\pi) \geqslant \Omega\left(\sigma\sqrt{kT}\right). \tag{2.2}$$

This result is agnostic w.r.t. the instance parameters, and can be obtained by optimizing for the worst possible ones. In such instances, we have Δ_i in the order of $\sqrt{\frac{k}{T}}$. It turns out that such a choice is the one balancing the cost of identifying a good action and the cost of making a mistake in the worst way.

Remark 2 (Failure of the Greedy Policy)**.** *It is easy to show that a greedy policy,* i.e.*, the policy* $\pi^G(t) \in \arg\max_{i\in[k]} \frac{1}{N_{i,t}} \sum_{l=1}^{t} X_{I_l,l} \mathbb{1}_{\{I_t=i\}}$ *that chooses the action* I_t *with the largest empirical mean so far, can suffer a regret linear in* T*. For instance, construct an instance where, with constant probability, the optimal action may yield a reward* $X_{i^*,t}$ *that is smaller than the minimum value of the support of* ν_i*, for some* $i \neq i^*$*. Then, with constant probability, this happens the very first time the optimal action is chosen and is then played never again, since the empirical mean of* i *can never go below* $X_{i^*,t}$*. Thus, we have* $R_{\boldsymbol{\nu},T}(\pi^G) \geqslant \Omega(T)$*. Intuitively, this is caused by the lack of an active exploration strategy from the policy. By definition, the greedy policy only tries to exploit the good actions so far, and is likely to miss the optimal action if that doesn't*

Algorithm 2: `UCB1` (Auer et al., 2002a)

1 Initialize $N_{i,1} \leftarrow 0$ for every $i \in [k]$ **for** $t \in [T]$ **do**
2 Compute $I_t \in \arg\max_{i \in [k]} \text{UCB}_{i,t} := \widehat{\mu}_{i,t} + \sigma\sqrt{\frac{4 \ln T}{N_{i,t}}}$
3 Play action I_t and observe $X_{i,t}$
4 Update $\widehat{\mu}_{i,t+1}$ for every $i \in [k]$
5 Update $N_{I_t,t+1} \leftarrow N_{I_t,t} + 1$
6 **end**

perform well from the beginning.

The Optimism Principle and UCB Policies Regret lower bounds set the bar on what a policy can hope to obtain in terms of suffered regret, and a natural question is what kind of policies are order-optimal in terms of regret upper bounds. The seminal work Auer et al. (2002a) introduces the *optimism principle* and the `UCB1` policy, a modification of the greedy policy that leads to exploration of *promising* actions (that's why the policy is called *optimistic*). Before describing the `UCB1` policy, we recall an important result on the convergence of mean estimators.

Proposition 2.1 (Upper Confidence Bound). *Let $\delta \in (0, 1)$. Let $\{X_t\}_{t=1}^n$ be a sequence of independent and σ-subgaussian random variable with mean μ, and $\widehat{\mu}_n := \frac{1}{n}\sum_{t=1}^n X_t$ is the empirical mean. Then, the following holds:*

$$\mathbb{P}\left(\widehat{\mu} + \sigma\sqrt{\frac{2 \ln \delta^{-1}}{n}} \geqslant \mu\right) \geqslant 1 - \delta. \tag{2.3}$$

We call $UCB_{i,t} := \widehat{\mu}_{i,t} + \sigma\sqrt{\frac{2 \ln \delta^{-1}}{N_{i,t}}}$ the *upper confidence bound* on the empirical mean $\widehat{\mu}_{i,t}$ of action i at round t (with the convention that $UCB_{i,t} = \infty$ if $N_{i,t} = 0$). Equation (2.3) gives us a way to upper bound with high probability the true mean of an action's reward distribution. This overestimation strategy allows a policy to be optimistic regarding the expected reward of a given action, and that the fundamental difference between the greedy policy and the `UCB1` policy, for which the algorithmic steps are reported in Algorithm 2. The `UCB1` algorithm runs as a greedy algorithm, but follows UCBs instead of the empirical means (line 2). The UCB is computed setting $\delta = T^{-2}$. This choice ensures that the width of the confidence interval is large enough to explore all actions by the end of the trial properly. Choosing the action with the largest UCB means considering the potential for an action, rather than the past performance alone. UCBs decrease with the number of pulls, and this helps in gaining information about the less explored actions. A small modification to this algorithm, that involves setting a dynamic $\delta_t^{-1} = t^{-2}$ makes `UCB1` an *anytime* algorithm, *i.e.*, an algorithm that is agnostic of T, and doesn't require its

knowledge (Auer and Ortner, 2010).

It is now time to state the instance-dependent regret upper bound of `UCB1`.

Theorem 2.3 (Instance-Dependent Regret Upper Bounds of `UCB1`)**.** *For every stochastic MAB instance $\boldsymbol{\nu}$, and time horizon T, the `UCB1` policy π^{UCB1} satisfies*

$$R_{\boldsymbol{\nu},T}(\pi^{UCB1}) \leqslant \mathcal{O}\left(\sum_{i\in[k]:\Delta_i>0} \sigma\left(\frac{\ln T}{\Delta_i} + \Delta_i\right)\right), \tag{2.4}$$

where the $\mathcal{O}$ notation only hides universal constants.

If we compare Equation (2.4) with Equation (2.1) we can see how the instance-dependent regret bound of `UCB1` is tight in T, up to universal constants, to best possible one, *i.e.*, the one of the lower bound.

Proof Sketch. To get an intuition on how this result is proven, we provide the fundamental steps of the proof (without delving too much into technical details).

Let $\delta = T^2$. We introduce a *good event* $\mathcal{G}$ defined as

$$\mathcal{G} = \left\{\mu_i \in \left[\widehat{\mu}_{i,t} \pm \sigma\sqrt{\frac{4\ln T}{N_{i,t}}}\right], \quad \forall i \in [k]\ \forall t \in [T]\right\}.$$

Under the good event $\mathcal{G}$, all actions have their means contained between the lower and the upper confidence bound in every round. A simple union bound yields $\mathbb{P}\left(\mathcal{G}^C\right) \leqslant \frac{1}{T^2}$.

Consider the following fundamental decomposition. For every policy π, it holds

$$R_{\boldsymbol{\nu},T}(\pi) = \sum_{i\in[k]} \mathbb{E}[N_{i,T}]\Delta_i.$$

Thus, bounding the number of times every suboptimal action $i \neq i^*$ is chosen is sufficient to get a bound on the regret.

The number of times a suboptimal action is chosen can be decomposed again, separating trials in which the good event $\mathcal{G}$ holds from the ones in which it doesn't.

$$\mathbb{E}\left[N_{i,T}\right] = \mathbb{E}\left[N_{i,T}\mathbb{1}_{\mathcal{G}}\right] + \mathbb{E}\left[N_{i,T}\mathbb{1}_{\mathcal{G}^C}\right] \leqslant \mathbb{E}\left[N_{i,T}\mathbb{1}_{\mathcal{G}}\right] + T\mathbb{P}\left(\mathcal{G}^C\right).$$

Thus, in expectation, the trials in which the good event doesn't hold only contribute in a constant way to the regret, and we can restrict our analysis to the trials in which $\mathcal{G}$ holds.

The proof can be concluded by bounding the summation of the confidence intervals for every suboptimal action $i \neq i^*$. It is possible to provide a bound of

$$\mathbb{E}[N_{i,t}] \leqslant \mathcal{O}\left(\frac{\ln T}{\Delta_i^2}\right).$$

Plugging this result into the previous decomposition yields the result.

This powerful arguments will be used to prove many of the results presented in this thesis.

It is worth noticing the dependence on Δ_i^{-2} of the number of times a suboptimal action is pulled. From a statistical perspective, this quantity be seen as a sample complexity of distinguishing a suboptimal action from the optimal one.

It is worth noticing that, when Δ_i is very small, the instance-dependent upper bound can become linear. In this case, we can provide a worst-case regret upper bound that shows that also in that case the regret suffered by `UCB1` is at most in the order of $\sqrt{T}$.

Theorem 2.4 (Instance-Dependent Regret Upper Bounds of `UCB1`)**.** *For every time horizon T, the `UCB1` policy π^{UCB1} satisfies*

$$\max_{\boldsymbol{\nu}} R_{\boldsymbol{\nu},T}(\pi^{UCB1}) \leqslant \mathcal{O}\left(\sqrt{kT \ln T}\right), \tag{2.5}$$

where the $\mathcal{O}$ notation only hides universal constants.

The bound presented in Equation (2.5) is tight to the corresponding lower bound in Equation (2.2) up to a $\sqrt{\ln T}$ term. In the bandits literature, it is customary to only look at the dominant term in the regret bound. However, it is possible to show that an algorithm called `MOSS` (Wei and Srivastava, 2020) has a tight minimax regret bound, getting rid of the logarithmic term.

2.2. Markov Decision Processes

In this section, we provide a quick overview of Markov Decision Processes (MDPs, Puterman (2014)) and their connection to MABs. This section is important in understanding the contributions presented in Section 3.

MDPs can be presented as generalization of MABs. In particular, they include the concept of *state* in the learning protocol. In MABs, we assumed that the set of available actions, as well as their reward-generation mechanisms, are the same at every round. In MDPs this is not true. At every round, the learner is faced with a different situation, called state, and it is the result of its previous actions (and possibly, the time step t, even though literature refers to this type of MDPs

Algorithm 3: Learner-Environment Interaction Protocol for Stochastic MDPs

1 The initial state $s_1 \sim \rho$ is sampled. **for** $t \in [T]$ **do**
2 Learner observes s_t and plays an action $I_t \in [k]$ (possibly at random)
3 Environment receives I_t and provides a reward X_t s.t. $\mathbb{E}[X_t \mid s_t, I_t] = f(s_t, a_t)$.
4 Learner receives $X_{I_t,t}$ and updates its internal status.
5 Environment provides a new state $s_{t+1} \sim P(\cdot \mid s_t, I_t)$.
6 **end**

as *non-stationary* MDPs).

This dimension brings several difficulties: the learner is now required to do *planning*, consider the future implications of performing a certain action in a given state, and not only focus on the immediate reward generated by an action. An additional difficulty is given by the fact that *transitions* may be stochastic, and the learner may need to estimate the transition probability from the current state to another.

MABs can be thought of as simpler MDPs where only one state exists, and every action makes the learner transition from the state to itself.

Setting Formalization

A finite MDP is a tuple $\mathcal{M} = (\mathcal{S}, [k], P, f, \rho)$, where $\mathcal{S}$ is a finite state space, $[k]$ is the action set [1], $P(\cdot \mid s, i)$ is the transition kernel, $f(s, i)$ is the reward function and ρ is the initial state probability distribution.

The interaction protocol between a learner and an MDP is formalized in Algorithm 3. The learner starts in state s_1, sampled from ρ (line 1). At each round $t \in T$, the learner observes s_t and chooses I_t (line 2), receives X_t sampled by the environment s.t. $\mathbb{E}[X_t \mid s_t, I_t] = f(s_t, a_t)$ (line 3), and the environment draws $s_{t+1} \sim P(\cdot \mid s_t, I_t)$ (line 5).

Remark 3 (MABs are special MDPs.)**.** *A stochastic MAB is the degenerate MDP with $|\mathcal{S}| = 1$ and identity transitions; the definitions above reduce to Section 2.1.*

Learning Goal

In MDPs, a (possibly randomized) policy π maps is defined as in MABs, with the difference that the history of interactions now also include the states encountered. The policy provides a decision based on the history and the current status, and we write $\pi(i \mid s)$. One of the learning goal that can be set in MDPs is the *average-reward regret*. Given an *unknown* MDP $\mathcal{M}$, the

[1] In this thesis, we will only consider MDPs with a fixed action set that is identical in every state, as in MABs

cumulative reward of a learner after T rounds can be defined in a similar way as in MABs:

$$J_{\mathcal{M},T}(\pi) := \mathbb{E}\left[\sum_{t\in[T]} r(s_t, I_t) \mid s_1 \sim \rho\right],$$

where the expectation is taken over the randomness of both the environment (rewards and transitions) and the policy/algorithm. This definition immediately implies the definition of optimal policy as in MABs, that we will call $\pi^*_{\mathcal{M},T}$, and its expected cumulative reward $J^*_{\mathcal{M},T}$. The learning goal is still regret minimization, and it can be defined as learning the policy π that minimizes the expected cumulative regret

$$R_{\mathcal{M},T}(\pi) = J^*_{\mathcal{M},T} - J_{\mathcal{M},T}(\pi).$$

In general, it is not possible to optimize this objective function in MDPs, due to the online nature of this metric and the complex structure that an MDP may have. However, there are special types of MDPs, such as *communicating* MDPs, where every state can reach every other under some policy. For this type of MDPs, stationary optimal policies exist in the *average-reward sense*. For a stationary π, let

$$\bar{J}_{\mathcal{M}}(\pi) \;=\; \lim_{T\to\infty} \frac{1}{T} J_{\mathcal{M},T}(\pi)$$

denote its *long-run average reward*, and $\rho^* = \max_\pi \rho(\pi)$. However, this metric is weaker than the standard *finite-time* notion of regret introduced beforehand, and is outside of the scope of this thesis.

3 | Between Bandits and Markov Decision Processes

This chapter focuses on online learning problems that are at the intersection of MABs and MDPs, where meaningful guarantees on finite-time regret can be provided. In general, MDPs cannot be tackled from the finite-time regret minimization perspective, because their flexible structure allows for the construction of MDPs where a single trial is not enough to obtain enough information. Indeed, in non-communicating MDPs, a single mistake may lead to an arbitrarily large portion that cannot be observed anymore. The MAB settings presented in this chapter can all be interpreted as special MDPs, for which providing algorithms with sub-linear finite-time expected regret is possible. The study of these special settings, which are of practical and theoretical interest, constitutes part of this thesis contribution. Some of them fall under the umbrella of *non-stationary* bandit settings, in which the state evolves only based on time. One can imagine such settings as *chain* MDPs where the state s_t is only composed of the time t and the transition is deterministic and independent of I_t. Moreover, it is worth keeping in mind that, in general, computing the optimal policy in a known MDP is an NP-hard problem, while in bandits it is a trivial scalar maximization. Nonetheless, we show that sub-linear finite-time regret can also be achieved in some settings in which the optimal policy is NP-hard to compute.

3.1. Autoregressive Bandits

Autoregressive processes naturally arise in a large variety of real-world scenarios, including stock markets, sales forecasting, weather prediction, advertising, and pricing. When facing a sequential decision-making problem in such a context, the temporal dependence between consecutive observations should be properly accounted for, guaranteeing convergence to the optimal policy. In this section, we propose a novel online learning setting, namely, Autoregressive Bandits (ARBs), in which the observed reward is governed by an autoregressive process of order m, whose parameters depend on the chosen action. We show that, under mild assumptions on the reward process, the optimal policy can be conveniently computed. Then, we devise a new optimistic regret minimization algorithm, namely, `AutoRegressive Upper Confidence`

Bound (AR−UCB), that suffers sublinear regret of order $\widetilde{\mathcal{O}}\left(\frac{(m+1)^{3/2}\sqrt{kT}}{(1-\Gamma)^2}\right)$, where T is the optimization horizon, k is the number of actions, and $\Gamma < 1$ is a stability index of the process. Finally, we empirically validate our algorithm, illustrating its advantages relative to bandit baselines and its robustness to the misspecification of key parameters.

This section presents Bacchiocchi et al. (2024), a joint project with Francesco Bacchiocchi, Marco Mussi, Davide Maran, Marcello Restelli, Nicola Gatti and Alberto Maria Metelli, published at the *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

3.1.1. Introduction

In a large variety of sequential decision-making problems, a learner is required to choose an action that, when executed, determines: (i) the immediate reward and (ii) the behavior of an underlying process that will influence, in some unknown manner, the future rewards. This process is influenced by the course of actions the agent performs and generates a temporal dependence between the sequence of observed rewards. A class of stochastic processes widely employed to model the temporal dependencies in real-world phenomena is the *autoregressive* (AR) processes (Hamilton, 2020). In this section, we model the reward of a sequential decision-making problem as an AR process whose parameters depend on the action selected by the agent at every round. This scenario can be represented as a particular class of *continuous* MDPs, where an AR process governs the temporal structure of the observed rewards through the action-dependent AR parameters that are unknown to the agent. It is worth mentioning that such a scenario displays notable differences compared to more traditional *non-stationary* learning problems. Indeed, in the scenario we address, the environment does not change, and the reward dynamics depend on the agent's course of actions only.

Original Contribution In this section we propose a novel setting, named *AutoRegressive Bandit* (ARB), in which the reward follows an AR process of order m whose parameters depend on the agent's actions. Importantly, we show that the optimal policy, differently from many bandit models, is *stationary* and *closed-loop*, as the optimal action depends on the previously observed rewards (Section 3.1.2). Then, we devise a new optimistic algorithm, namely AutoRegressive Upper Confidence Bound (AR−UCB), to learn an optimal policy in an online fashion (Section 3.1.4), and we show that it suffers sublinear regret of order $\widetilde{\mathcal{O}}\left(\frac{(m+1)^{3/2}\sqrt{kT}}{(1-\Gamma)^2}\right)$, where T is the optimization horizon, k is the number of actions, and $\Gamma < 1$ is a stability index of the process (Section 3.1.5). Finally, we empirically evaluate AR−UCB comparing its performance with several bandit baselines with competitive results and illustrating its notable robustness w.r.t. the misspecification of key parameters (Section 3.1.6).

3.1.2. Problem Formulation

In this section, we introduce the ARB setting, formalize the learning problem, how the learner interacts with the environment, assumptions, policies and definition of regret (Section 3.1.2). Subsequently, we derive a closed-form solution for the optimal policy of an ARB (Section 3.1.2).

Setting

We study the sequential interaction between a learner and an environment. At every round $t \in [T]$, the learner chooses an action $I_t \in [k]$. In the ARB setting, the reward evolves according to an *autoregressive process* of order m (AR(m), Hamilton, 2020). Thus, the learner observes a random reward X_t of the form:

$$X_t = \gamma_0(I_t) + \sum_{i=1}^{m} \gamma_i(I_t) X_{t-i} + \xi_t, \tag{3.1}$$

where $X_t \in \mathcal{X}$ ($\mathcal{X} \subseteq \mathbb{R}$ is the reward space), $\gamma_0(I_t) \in \mathbb{R}$ and $(\gamma_i(I_t))_{i\in[m]} \in \mathbb{R}^m$ are the unknown *parameters* depending on chosen action I_t, and ξ_t is a zero-mean σ^2-subgaussian random noise, independent conditioned to the past. The reward evolution can be expressed in an alternative form as follows:[1]

$$X_t = \langle \boldsymbol{\gamma}(I_t), \mathbf{Z}_{t-1} \rangle + \xi_t, \tag{3.2}$$

where $\mathbf{Z}_{t-1} := (1, X_{t-1}, \dots, X_{t-m})^T \in \mathcal{Z} := \{1\} \times \mathcal{X}^m$ is the *vector of past rewards* expressing past history, and $\boldsymbol{\gamma}(i) := (\gamma_0(i), \dots, \gamma_m(i))^T \in \mathbb{R}^{m+1}$ is the *parameter vector*, defined for all the actions $i \in [k]$. It is worth noting that when $\gamma_j(i) = 0$ for all $j \in [m]$ and $i \in [k]$, the ARB setting reduces to a standard MAB (Auer et al., 2002a).

Assumptions We introduce the assumption that we employ in the section and comment on its role.

Assumption 1. *The parameters $\boldsymbol{\gamma}(i)$ fulfill the following conditions:*

a. (Non-negative coefficients) $\gamma_j(i) \geqslant 0$ for every $i \in [k]$, $j \geqslant 0$;
b. (Stability) $\Gamma := \max_{i\in[k]} \sum_{j=1}^{m} \gamma_j(i) < 1$;
c. (Boundedness) $g := \max_{i\in[k]} \gamma_0(i) < +\infty$.

First, Assumption 1.a requires that the coefficients of the AR process are non-negative. This scenario is ubiquitous in real-world AR phenomena (*e.g.*, pricing, stock markets, digital adver-

[1] Although the linear structure might resemble the *contextual linear bandits* (Chu et al., 2011), the two settings are non-comparable. Indeed, in our ARBs the vector $\mathbf{Z}_{t-1}$ is not sampled independently at every round, but, instead, follows a sequential process depending on the past, making the decision problem way more challenging.

tising), where processes violating such an assumption will generate unrealistic sign alternation behaviours. An extensive discussion and a graphical elaboration about this assumption are provided in Appendix A.3. Assumption 1.b requires that the sum of $\boldsymbol{\gamma}(i)$ is limited to a value $\Gamma \in [0, 1)$ and Assumption 1.c enforces the boundedness of $\gamma_0(i)$, for every $i \in [k]$. These latter assumptions ensure that the AR process does not diverge in expectation, regardless of the sequence of actions played.

Connection to MDPs

The ARB interaction protocol can be seen as a special class of MDPs. In particular, we consider the class of *continuous state* MDPs, a variation of the finite MDP introduced in Section 2.2 where the state space $\mathcal{S}$ is continuous. Let $s_t = \mathbf{Z}_{t-1}$ be the state at time t. Then, the learner can observe the state s_t before providing a decision I_t, and the reward function is defined as $f(s_t, I_t) = \langle \boldsymbol{\gamma}(I_t), s_t \rangle$, and the actual reward is $X_t = f(s_t, I_t) + \xi_t$. The transition is stochastic and depends on the reward, and yields s_{t+1} by shifting the elements of s_t by one space to the right (with the exclusion of the first element) and inserting X_t. Finally, the initial state is provided deterministically as $\mathbf{Z}_0$.

Learning Goal

We formally define a policy in the ARB setting, and we provide the learning goal in this setting. These definitions are coherent with the ones proposed in Chapter 2, but specifically casted on the ARB problem. The learner's behavior is modeled by a deterministic policy $\boldsymbol{\pi} = (\pi_t)_{t \in \mathbb{N}}$ defined, for every round $t \in \mathbb{N}$ as $\pi_t : \mathcal{H}_{t-1} \to \mathcal{A}$, mapping the history of observations $H_{t-1} = (x_0, a_1, x_1, \ldots, a_{t-1}, x_{t-1}) \in \mathcal{H}_{t-1}$ to an action $I_t = \pi_t(H_{t-1}) \in \mathcal{A}$ where $\mathcal{H}_{t-1} = \mathcal{X} \times (\mathcal{A} \times \mathcal{X})^{t-1}$ is the set of histories of length $t-1$. The performance of a policy $\boldsymbol{\pi}$ is evaluated in terms of the *expected cumulative reward* over the horizon $T \in \mathbb{N}$, defined as:

$$J_T(\boldsymbol{\pi}) := \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \tag{3.3}$$

with:

$$X_t = \langle \boldsymbol{\gamma}(I_t), \mathbf{Z}_{t-1} \rangle + \xi_t,$$
$$I_t = \pi_t(H_{t-1}),$$

where the expectation is taken w.r.t. the randomness of the reward noise ξ_t. A policy $\boldsymbol{\pi}^*$ is *optimal* if it maximizes the expected average reward, *i.e.*, $\boldsymbol{\pi}^* \in \arg\max_{\boldsymbol{\pi}} J_T(\boldsymbol{\pi})$, whose performance is denoted as $J_T^* := J_T(\boldsymbol{\pi}^*)$. The goal of the learner is to minimize the *expected cumulative (policy) regret* by playing a policy $\boldsymbol{\pi}$, competing against the optimal policy $\boldsymbol{\pi}^*$ over a *learning*

horizon $T \in \mathbb{N}^+$:

$$R(\boldsymbol{\pi}, T) = J_T^* - J_T(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right], \tag{3.4}$$

where $r_t := X_t^* - X_t$ is the instantaneous policy regret and $(X_t^*)_{t \in [\![T]\!]}$ is the sequence of rewards observed by playing the optimal policy $\boldsymbol{\pi}^*$.

Optimal Policy

In this section, we derive a closed-form expression for the optimal policy $\boldsymbol{\pi}^*$ for the expected cumulative reward, under Assumption 1.a.

Theorem 1 (Optimal Policy)**.** *Under Assumption 1.a, for every round* $t \in [T]$*, the optimal policy* $\pi_t^*(H_{t-1})$ *satisfies:*

$$\pi_t^*(H_{t-1}) \in \arg\max_{i \in [k]} \langle \boldsymbol{\gamma}(i), \mathbf{Z}_{t-1} \rangle. \tag{3.5}$$

This result deserves some comments. First, the optimal action depends on the vector of past rewards $\mathbf{Z}_{t-1}$ and, thus, on the most recent m rewards $x_{t-1}, \dots, x_{t-m}$ only. Thus, the optimal policy $\boldsymbol{\pi}^*$ is non-Markovian with memory m or, equivalently, Markovian w.r.t. the state representation $\mathbf{Z}_{t-1}$.[2] Second, the optimal action maximizes, at every round $t \in [T]$, the *expected instantaneous reward* $\mathbb{E}[X_t | H_{t-1}] = \langle \boldsymbol{\gamma}(i), \mathbf{Z}_{t-1} \rangle$. This is a consequence of the non-negativity of the parameters $\gamma_j(i)$ (Assumption 1.a), which enforces a meaningful evolution of the AR process, compatible with our real-world motivating scenarios. This way, the action maximizing the expected *immediate* reward (*i.e.*, a *myopic* policy) is optimal for the expected *cumulative* reward too. The proof can be found in Appendix A.1.

3.1.3. Related Works

In this section, we discuss and compare the works that share similarities with the Autoregressive Bandits. We analyze both solutions related to multi-armed bandits and online learning in non-linear systems.

Multi-Armed Bandits In the more classical Multi-Armed Bandit (MAB) setting, the learning problem does not involve temporal dependencies between rewards. The MAB setting has been studied under the assumptions of both *stochastic* and *adversarial* noise models. In the former case, `UCB1` (Lai and Robbins, 1985; Auer et al., 2002a) represents the parent algorithm. Instead, when adversarial noise is involved `EXP3` (Auer et al., 1995, 2002b) is usually employed. This

[2] We can look at the ARB as a particular *Markov Decision Processes* (MDPs, Puterman, 2014) with $\mathbf{Z}_{t-1} \in \mathcal{Z}$ as state representation.

algorithm has been extended by REXP3 (Besbes et al., 2014) to handle with the *non-stationary* setting. Differently from both the adversarial and non-stochastic setting, we assume that the rewards are not preselected by an adversary or nature but, instead, they change as an effect of the actions played. Indeed, the underlying autoregressive process (affected by a stochastic noise) is such that the current action impacts the future rewards. Therefore, importing the adversarial MAB terminology, the ARBs can be reduced to an adversary setting with an *adaptive* (or non-oblivious) adversary (Dekel et al., 2012b). In particular, the $\widetilde{\mathcal{O}}(\sqrt{kT})$ regret guarantees of EXP3 are not achievable in the ARB setting as EXP3 competes against the best constant policy while the optimal policy for ARBs is not constant (see Theorem 1 and Section 3.1.6). Moreover, our setting presents similarities with MABs with *delayed* feedback (*e.g.*, Pike-Burke et al., 2018). However, in ARB the effect of the actions is propagated (not exactly delayed). Markov (Ortner et al., 2012) and restless (Tekin and Liu, 2012) bandits, instead, consider underlying processes that influence the rewards. However, these processes are not supposed to be controlled by the action history. Other works (*e.g.*, Mussi et al., 2023) consider complex action-dependent feedback vanishing over time. In Chen et al. (2023), the authors study the control problem in a setting that considers temporal structure modeled as an AR(1) process.

Online Learning in Non-Linear Systems The ARB setting is a specific case of a non-linear dynamical system. Although the literature related to this setting is wide, no work faces all problems that the ARB setting presents, including learning to control with regret guarantees. Mania et al. (2022) focus on learning the parameters of a particular class of non-linear systems. However, the approach is limited to estimation and no control algorithm is proposed. Similarly, Umlauft and Hirche (2017) deal with learning the system parameters with stability guarantees without the chance to control it. Several recent works (*e.g.*, Kakade et al., 2020; Lale et al., 2021) focus on the learning and control of non-linear systems with regret guarantees. However, these works make use of an oracle to solve a complex optimization problem to perform optimistic planning (*i.e.*, optimal policy given an optimistic estimate of the system). This problem in a non-linear setting, however, is proven to be NP-hard (Sahni, 1974; Dani et al., 2008). Furthermore, the class of non-linear systems considered in these works does not include the ARB setting. Other works (*e.g.*, Albalawi et al., 2021) overcome the request for the oracle by searching in the restricted space of constant policies, leading to the best equilibrium. However, this solution can be suboptimal in several cases, including ARBs (see Section 3.1.6).

3.1.4. Autoregressive Upper Confidence Bound

In this section, we present AutoRegressive Upper Confidence Bound (AR-UCB), an optimistic regret minimization algorithm for the ARB setting whose pseudo-code is reported in Algorithm 4. AR-UCB leverages the myopic optimal policy for ARBs (Theorem 1) and imple-

Algorithm 4: AR−UCB.

Input : Regularization parameter $\lambda > 0$, autoregressive order m

1 Initialize $\mathbf{V}_0(i) = \lambda \mathbf{I}_{k+1}$, $\mathbf{b}_0(i) = \mathbf{0}_{k+1}$, $\widehat{\boldsymbol{\gamma}}_0(i) = \mathbf{0}_{k+1}$, $\forall i \in [k]$, $\mathbf{Z}_0 = (1, 0, \dots, 0)^T$
2 **for** $t \in [T]$ **do**
3 Compute $I_t \in \arg\max_{i \in [k]} \text{UCB}_{i,t} := \langle \widehat{\boldsymbol{\gamma}}_{t-1}(i), \mathbf{Z}_{t-1} \rangle + \beta_{t-1}(i) \, \|\mathbf{Z}_{t-1}\|_{\mathbf{V}_{t-1}(i)^{-1}}$
4 Play action I_t and observe $X_t = \langle \boldsymbol{\gamma}(I_t), \mathbf{Z}_{t-1} \rangle + \xi_t$
5 **for** $i \in [k]$ **do**
6 $\mathbf{V}_t(i) = \mathbf{V}_{t-1}(i) + \mathbf{Z}_{t-1} \mathbf{Z}_{t-1}^T \mathbb{1}_{\{i = I_t\}}$
7 $\mathbf{b}_t(i) = \mathbf{b}_{t-1}(i) + \mathbf{Z}_{t-1} X_t \mathbb{1}_{\{i = I_t\}}$
8 $\widehat{\boldsymbol{\gamma}}_t(i) = \mathbf{V}_t(i)^{-1} \mathbf{b}_t(i)$
9 **end**
10 Update $\mathbf{Z}_t = (1, X_t, \dots, X_{t-k+1})^T$
11 **end**

ments an incremental regularized least squares procedure to estimate the unknown parameters $\boldsymbol{\gamma}(i)$, for every action $i \in [k]$ independently. The algorithm requires the knowledge of the order m of the AR process, although this knowledge can be replaced with the one of an upper bound $\overline{m} > m$ of the AR order.[3]

AR−UCB starts by initializing for all the actions $i \in [k]$ the Gram matrix $\mathbf{V}_0(i) = \lambda \mathbf{I}_{k+1}$, where $\lambda > 0$ is the Ridge regularization parameter, the vectors $\mathbf{b}_0(i) = \widehat{\boldsymbol{\gamma}}_0(i) = \mathbf{0}_{k+1}$, and the observations vector $\mathbf{z}_0 = (1, 0, \dots, 0)^T$ (line 1).[4] Then, for each round $t \in [T]$, AR−UCB computes the *Upper Confidence Bound* (UCB) index (line 3) for every $i \in [k]$. Such an optimistic index is composed of the inner product between the estimated value of $\boldsymbol{\gamma}(i)$ and the state representation $\mathbf{Z}_{t-1}$, plus the confidence interval $\beta_{t-1}(i)$. Formally:

$$I_t \in \arg\max_{i \in [k]} \text{UCB}_{i,t} := \langle \widehat{\boldsymbol{\gamma}}_{t-1}(i), \mathbf{Z}_{t-1} \rangle + \beta_{t-1}(i) \, \|\mathbf{Z}_{t-1}(i)\|_{\mathbf{V}_{t-1}(i)^{-1}} , \tag{3.6}$$

where $\widehat{\boldsymbol{\gamma}}_{t-1}(i)$ is the most recent estimate of the parameter vector $\boldsymbol{\gamma}(i)$, $\mathbf{Z}_{t-1} = (1, x_{t-1}, \dots, x_{t-k})^T$ is the observations vector, and $\beta_{t-1}(i) \geqslant 0$ is an exploration coefficient that will be defined later (Section 3.1.5). The index $\text{UCB}_{i,t}$ is designed to be optimistic, *i.e.*, $\langle \boldsymbol{\gamma}(i), \mathbf{Z}_{t-1} \rangle \leqslant \text{UCB}_{i,t}$ with high probability for all $i \in [k]$. Then, action I_t is executed (line 4) and the new reward X_t is observed. This sample is employed to update the Gram matrix estimate $\mathbf{V}_t(I_t)$, the vector $\mathbf{b}_t(I_t)$, and the estimate $\widehat{\boldsymbol{\gamma}}_t(I_t)$ (line 8).

[3] Indeed, any AR process of order m can be regarded as an AR process of order $\overline{m} > m$ setting $\gamma_j(i) = 0$ for $j \in \{m+1, \dots, \overline{m}\}$. An empirical validation of the AR−UCB performances in the case of a misspecified m is provided in Section 3.1.6.

[4] We assume to know the initial observations vector $\mathbf{z}_0$. If this is not the case, we can play an arbitrary action for the first k rounds to observe $(x_t)_{t \in [k]}$ with just an additional constant loss term.

3.1.5. Regret Analysis

In this section, we present the analysis of the regret of AR−UCB. We start providing a self-normalized concentration inequality for estimating the AR parameters $\boldsymbol{\gamma}(i)$ (Section 3.1.5). Then, we derive a decomposition of the regret (Section 3.1.5) that is useful to complete the analysis and, finally, we present the bound on the expected cumulative (policy) regret (Section 3.1.5). The complete proofs of the theorems stated in this section can be found in Appendix A.1.

Concentration Inequality for the Parameter Vectors

We start by providing a concentration result for the estimates $\widehat{\boldsymbol{\gamma}}_t(i)$ of the true parameter vector $\boldsymbol{\gamma}(i)$, for every action $i \in [k]$, as performed in Algorithm 4. At the end of each round t, where the chosen action is $I_t \in [k]$, we solve the Ridge-regularized linear regression problem and update the coefficient vector estimate $\widehat{\boldsymbol{\gamma}}_t(I_t)$ associated with I_t:

$$\begin{aligned}\widehat{\boldsymbol{\gamma}}_t(I_t) &= \operatorname*{arg\,min}_{\widetilde{\boldsymbol{\gamma}} \in \mathbb{R}^{k+1}} \sum_{l \in \mathcal{O}_t(I_t)} (X_l - \langle \widetilde{\boldsymbol{\gamma}}, \mathbf{Z}_{l-1} \rangle)^2 + \lambda \left\| \widetilde{\boldsymbol{\gamma}} \right\|_2^2 \\ &= \mathbf{V}_t(I_t)^{-1} \mathbf{b}_t(I_t),\end{aligned}$$

where $\mathcal{O}_t(i)$ is the set of rounds where action i has been chosen, *i.e.*, $\mathcal{O}_t(i) \coloneqq \{\tau \in [t] : I_\tau = i\}$. The following result shows how the estimate $\widehat{\boldsymbol{\gamma}}(a)$ concentrates around the true parameters $\boldsymbol{\gamma}(a)$ over the rounds.

Lemma 2 (Self-Normalized Concentration)**.** *Let $i \in [k]$ be an action, let $\{\widehat{\boldsymbol{\gamma}}_t(i)\}_{t \in \mathcal{O}_\infty(i)}$ be the sequence of solutions to the Ridge regression problems computed by Algorithm 4. Then, for every regularization parameter $\lambda > 0$, confidence $\delta \in (0, 1)$, simultaneously for every round $t \in [T]$ and action $i \in [k]$, with probability at least $1 - \delta$ it holds that:*

$$\left\| \widehat{\boldsymbol{\gamma}}_t(i) - \boldsymbol{\gamma}(i) \right\|_{\mathbf{V}_{t(i)}} \leqslant \sqrt{\lambda} \|\boldsymbol{\gamma}(i)\|_2 + \sigma \sqrt{2 \log\left(\frac{n}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(i)}{\lambda^{k+1}}\right)}.$$

Lemma 2 resembles the self-normalized concentration inequality of (Abbasi-Yadkori et al., 2011, Theorem 1). However, contrary to LIN−UCB (Abbasi-Yadkori et al., 2011), the exploration coefficients $\beta_t(i)$ are different for every action $i \in [k]$. Lemma 2 allows properly defining the exploration coefficients $\beta_t(i)$ employed in Algorithm 4, defined for every action $i \in [k]$ and round $t \in [T-1]$:

$$\beta_t(i) \coloneqq \sqrt{\lambda(g^2+1)} + \sigma \sqrt{2 \log\left(\frac{k}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(i)}{\lambda^{g+1}}\right)}. \tag{3.7}$$

This formula contains two terms. The first one is a *bias* term that increases with g (*i.e.*, the maximum value of the largest $\gamma_0(i)$ over the actions $i \in [k]$, see Assumption 1.c) and with the regularization parameter of the Ridge regression $\lambda > 0$. The second one is the *concentration* term and increases with the subgaussian parameter σ of the noise, the number of actions n, and the determinant of the design matrix $\mathbf{V}_t(i)$, but decreases in λ. It is worth noting that $\beta_t(i)$ is obtained from Lemma 2, by observing that, under Assumptions 1.b and 1.c, we have $\|\boldsymbol{\gamma}(i)\|_2 \leqslant \sqrt{g^2 + \Gamma^2} \leqslant \sqrt{g^2 + 1}$. Thus, the exploration coefficient $\beta_t(i)$ ensures that, with probability $1 - \delta$, the following inequality holds simultaneously for all actions $i \in [k]$ and rounds $t \in [T - 1]$:

$$\|\widehat{\boldsymbol{\gamma}}_t(i) - \boldsymbol{\gamma}(i)\|_{\mathbf{V}_t(i)} \leqslant \beta_t(i). \tag{3.8}$$

We observe that $\beta_t(i)$ (see Equation 3.7) and AR-UCB do not require the knowledge of the maximum sum Γ of the parameters $\gamma_j(i)$ over the actions (see Assumption 1.b). This is a desirable feature of our algorithm as Γ is often unknown in practice and difficult to upper bound or estimate. Nevertheless, Γ appears in the regret analysis in Section 3.1.5. Differently, the value of g, needed to compute the optimistic coefficient $\beta_t(i)$, can be easily replaced with an upper bound $\overline{g} > g$ when unknown.[5]

Regret Decomposition

In this section, we present a novel *decomposition* of the regret that will be employed in the final bound of Section 3.1.5. The contents of this section are of independent interest and applicable to any learner's policy $\boldsymbol{\pi}$, beyond AR-UCB. From a technical perspective, the analysis is composed of two steps: (i) we decompose the instantaneous (policy) regret r_t in terms of the instantaneous *external regret* ρ_t (Lemma 3); (ii) we bound the cumulative expected (policy) regret $R(\boldsymbol{\pi}, T) = \mathbb{E}[\sum_{t=1}^{T} r_t]$ in terms of the expected cumulative external regret $\varrho(\boldsymbol{\pi}, T) = \mathbb{E}[\sum_{t=1}^{T} \rho_t]$ (Lemma 4).

We start with step (i), by recalling that the definition of cumulative expected (policy) regret $R(\boldsymbol{\pi}, T)$ in Equation 3.4 compares the sequence of rewards $(X_t^*)_{\{t \in [T]\}}$ when executing the optimal policy $\boldsymbol{\pi}^*$ with the sequence of rewards $\{X_t\}_{\{t \in [T]\}}$ when executing the learner's policy $\boldsymbol{\pi}$. However, in our ARB setting, the observed reward X_t depends on the past history H_{t-1}. Thus, the instantaneous (policy) regret $r_t := X_t^* - X_t$ can be decomposed in two terms: (a) the dissimilarity between the past history H_{t-1}^* when executing the optimal policy and the learner's observed history H_{t-1}; (b) the instantaneous *external regret* (Dekel et al., 2012b) $\rho_t := \langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(I_t), \mathbf{Z}_{t-1} \rangle$ representing the loss of executing the learner action I_t instead of the optimal one $i_t^* = \pi_t^*(H_{t-1}^*)$ *assuming* that such actions are applied to the observations vector

[5] An empirical analysis of the effect of the misspecification of such a parameter is provided in Section 3.1.6.

$\mathbf{Z}_{t-1}$ generated by the execution of the learner's policy. The following result formalizes the instantaneous regret decomposition.

Lemma 3 (Policy Regret Decomposition)**.** *Let $(x_t^*)_{t\in[T]}$ be the sequence of rewards by executing the optimal policy $\boldsymbol{\pi}^*$ and let $(X_t)_{t\in[T]}$ be the sequence of rewards by executing the learner's policy $\boldsymbol{\pi}$. Then, for every $t \in [T]$ it holds that:*

$$\begin{aligned} r_t &= X_t^* - X_t \\ &= \sum_{j=1}^{m} \gamma_j(i_t^*)(X_{t-i}^* - X_{t-i}) + \langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(I_t), \mathbf{Z}_{t-1} \rangle \\ &= \sum_{j=1}^{m} \gamma_j(i_t^*) r_{t-i} + \rho_t, \end{aligned} \tag{3.9}$$

where $r_t := X_t^* - X_t$ is the instantaneous policy regret, $\rho_t := \langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{Z}_{t-1} \rangle$ is the instantaneous external regret, $i_t^* = \pi_t^*(H_{t-1}^*)$, and $r_{t-l} = 0$ if $l \geqslant t$.

The decomposition in Equation (3.9) comprises two terms. The second one ρ_t is the instantaneous external regret discussed above. The first one defines a recurrence relation of order k on the instantaneous policy regret r_t. We now move to step (ii) with the following result that shows that the contribution of the recurrence can be reduced to a term depending on Γ and k that multiplies the cumulative external regret.

Lemma 4 (External-to-Policy Regret Bound)**.** *Let $\boldsymbol{\pi}$ be the learner's policy and $T \in \mathbb{N}$ be the horizon. Under Assumptions 1.a and 1.b, it holds that:*

$$\begin{aligned} R(\boldsymbol{\pi}, T) &= \mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{j=1}^{m} \gamma_j(i_t^*) r_{t-j} + \rho_t\right]\right] \\ &\leqslant \left(\frac{\Gamma m}{1-\Gamma} + 1\right) \varrho(\boldsymbol{\pi}, T), \end{aligned} \tag{3.10}$$

where $\varrho(\boldsymbol{\pi}, T) := \mathbb{E}\left[\sum_{t=1}^{T} \rho_t\right]$ is the cumulative expected external regret.

Lemma 4 provide us a bound on the cumulative expected (policy) regret $R(\boldsymbol{\pi}, T)$ achieved by `AR-UCB` (or any algorithm playing in an ARB) by bounding the cumulative expected external regret $\varrho(\boldsymbol{\pi}, T)$. The order of the regret bound w.r.t. T is governed by the external regret, while the effect of a *weaker* history (*i.e.*, the sub-optimal actions of the past) emerges as an instance-specific constant. Such a constant is 1 whenever $m = 0$ or $\Gamma = 0$, *i.e.*, when the ARB reduces to a standard MAB. In all other cases, the bigger the value of m or Γ, the more visible the AR effects are, and, consequently, the more the sub-optimal choices of the past get amplified. Finally,

we point out that the multiplicative factor $\frac{\Gamma m}{1-\Gamma} + 1$ to pass from external to policy regret is tight since there exists a sequence of external regrets in which the inequality of Lemma 4 holds with equality (see Appendix A.1).

Regret Bound

In the following, we present a bound on the expected policy regret bound for AR−UCB.

Theorem 5. *Let $\delta = (2T)^{-1}$. Under Assumptions 1.a, 1.b, and 1.c,* AR−UCB *suffers a cumulative expected (policy) regret bounded by (highlighting the dependence on g, σ, m, Γ, k, and T):*

$$\mathbb{E}[R(\texttt{AR-UCB}, T)] \leqslant \widetilde{\mathcal{O}}\left(\frac{(g+\sigma)(m+1)^{3/2}\sqrt{kT}}{(1-\Gamma)^2}\right).$$

Some observations are in order. First, when we set $m = 0$ and $\Gamma = 0$, *i.e.*, we reduce the ARB to a standard MAB, we obtain a regret rate of $\widetilde{\mathcal{O}}((g+\sigma)\sqrt{kT})$, which is tight for standard MABs. The quantity $\frac{g+\sigma}{1-\Gamma}$ is the maximum value that rewards can achieve, as proven in Lemma 33. As intuition suggests, the ARB learning problem becomes more challenging as the AR order m increases and when the bound on the sum of the parameters Γ approaches one. This is witnessed in Theorem 5 with the dependence of the regret on $(m+1)^{3/2}$ and $(1-\Gamma)^{-1}$. The interplay between m and $(1-\Gamma)^{-1}$ shows that even if two instances have the same sum of parameters (*i.e.*, Γ), the one with fewer coefficients (*i.e.*, m) is more easily learnable. This is explained by the fact that our algorithm learns the individual parameters by means of a regression procedure learning to a $\sqrt{m+1}$ in the regret. Finally, suppose we run AR−UCB with a larger AR order $\overline{m} > m$. In such a case, the dependence on $(m+1)^{3/2}$ becomes $(m+1)(\overline{m}+1)^{1/2}$, since the factor due to passing from external to policy regret (Lemma 4) will always contain the true m, while $\overline{m}$ appears because of the estimation process. Similarly, if we execute AR−UCB with a value $\overline{g} > g$, the regret bound still holds by replacing g with $\overline{g}$.

Remark 4 (Comparison with Existing Results from MDPs Literature)**.** *If we consider our problem as an MDP, we are in an undiscounted finite-time scenario. This scenario is more challenging w.r.t. the episodic one. The ARB setting is not tabular (as it has continuous space) nor an LQR (as it has discrete actions). Our setting can be viewed as an Hölder continuous MDP by making a one-hot encoding of the k actions, but the regret bounds for this family of processes are, in the best-case scenario, in the order of $\widetilde{\mathcal{O}}\left(T^{2/3}\right)$, much worse than our bound of order $\widetilde{\mathcal{O}}\left(\sqrt{T}\right)$.*

3.1.6. Numerical Validation

In this section, we first provide (Section 3.1.6) a numerical validation of AR−UCB compared with other bandit baselines in synthetically-generated domains. Then, we discuss (Section 3.1.6) the importance of exploiting the noise in this setting, and, subsequently, we analyze the sensitivity of AR−UCB to the misspecification of the two most important parameters, *i.e.*, g (Section 3.1.6) and m (Section 3.1.6). Additional experimental results are provided in Appendix A.4. The code to reproduce the experiments can be found at `https://github.com/gianmarcogenalti/autoregressive-bandits`.

Running Time The algorithms are implemented in Python 3.11, and run over an Intel Core $i7-8750H$ @ 2.20 GHz with 16 GB DDR4 RAM. All the presented experiments took ≈ 10 minutes for a complete run.

AR−UCB vs Bandit Baselines

Setting We evaluate AR−UCB in three scenarios that differ in the properties of the autoregressive processes that govern the rewards. The competing algorithms are evaluated in terms of cumulative regret w.r.t. the setting-specific clairvoyant. The three settings have their AR(m) process order $m \in \{2, 4\}$, number of actions $k \in \{2, 7\}$, and scale $g \in \{1, 20, 920\}$. The values of $\boldsymbol{\gamma}(i)$ have been sampled from uniform probability distributions for each action $i \in [k]$ and for each setting. The environments are noisy with a standard deviation $\sigma \in \{0.75, 1.5, 10\}$. We chose to set the hyper-parameters of AR−UCB as follows: $\lambda = 1$, while $\overline{g} \in \{10, 100, 1000\}$, that is equivalent to chose $\overline{g}$ of the same magnitude of the true value g, in a pessimistic fashion. Table 3.1 summarizes the details of the three environments.

Baselines AR−UCB will compete with several bandit baselines. First, it is compared with UCB1 (Auer et al., 2002a), a widely adopted solution for stochastic MABs. Second, we consider EXP3, designed for adversarial MABs (Auer et al., 1995, 2002b) and its extension to finite-memory adaptive adversaries B−EXP3 (Dekel et al., 2012b). Lastly, we compare AR−UCB with AR2 (Chen et al., 2023), an algorithm for managing AR(1) processes. The hyper-parameters chosen for the baselines are the ones proposed in the original papers.

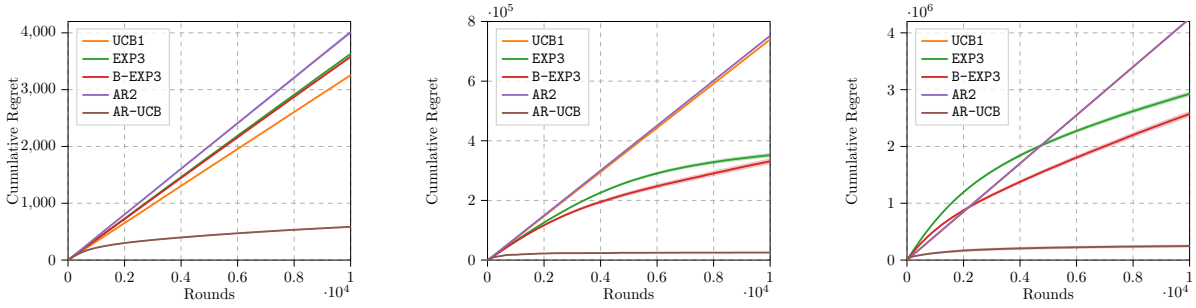
Results Figure 3.1 shows the average cumulative regrets for AR−UCB and the other bandit baselines. We observe that AR−UCB suffers the smallest cumulative regret in these scenarios, always displaying a sublinear behavior. Both EXP3 and B−EXP3 in two scenarios out of three (B and C) achieve sublinear regret. On the other hand, both UCB1 and AR2 are not able to achieve sublinear regret in the presented scenarios. This is not surprising since we require them to learn more complex processes than those they are designed for (*i.e.*, models with $m = 0$ and

Setting	Parameters m	k	g	σ
A	2	2	1	0.75
B	4	7	20	1.5
C	4	7	920	10

Table 3.1: Settings description.

σ	Stochastic	Deterministic
0	**19994 (0)**	**19994 (0)**
0.1	**20167 (0.20)**	19998 (2.04)
0.5	**22049 (1.02)**	20012 (1.02)
1.0	**24504 (2.04)**	20030 (2.04)
2.0	**29428 (4.09)**	20067 (4.08)

Table 3.2: Cumulative reward of the Stochastic and Deterministic clairvoyants (100 runs, mean (std)).



(a) Setting A. (b) Setting B. (c) Setting C.

Figure 3.1: Settings and cumulative regret of AR-UCB and multiple baselines (100 runs, mean $\pm$ std).

$m = 1$ for UCB1 and AR2, respectively).

On the Effect of Stochasticity

The optimal policy (Theorem 1) for the ARB setting exploits the contribution of the noise to increase the collected reward. In this section, we provide experimental evidence of this phenomenon. We first introduce a notion of *optimal policy without noise*. Then, we conduct an experiment to highlight the variations between the two policies in environments presenting different noise magnitudes.

Optimal Policy without Noise The optimal policy, when no noise is involved, is *constant* and corresponds, for sufficiently large T, to playing the action $i^+ \in [k]$ that brings the system to the most profitable steady state.[6] Such an action i^+ is the one maximizing the *steady-state reward*,

[6]The request for large T is to make transient effects neglectable.

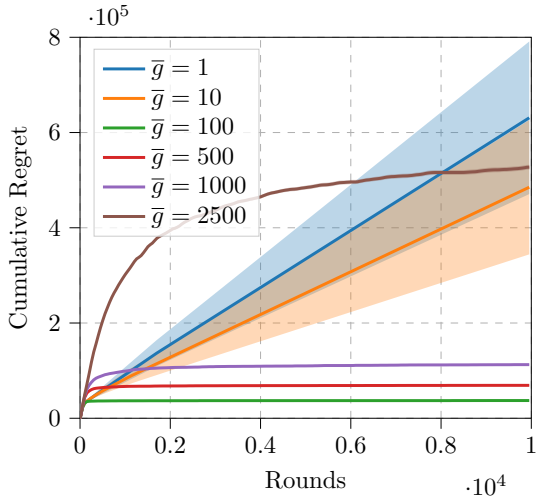


Figure 3.2: Effect of the choice of parameter $\overline{g}$ on the `AR-UCB` cumulative regret (100 runs, mean $\pm$ std).

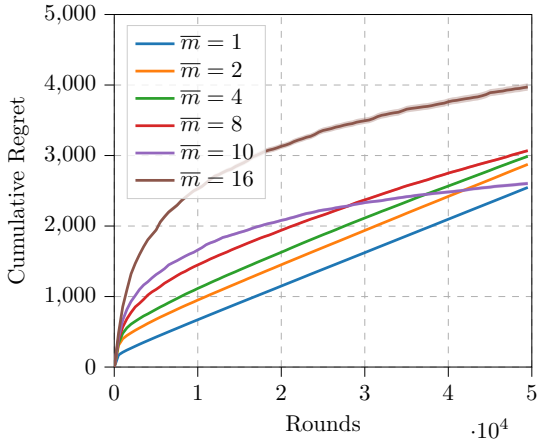


Figure 3.3: Effect of the choice of parameter $\overline{k}$ on the `AR-UCB` cumulative regret (100 runs, mean $\pm$ std).

namely:

$$i^+ \in \arg\max_{i \in [k]} \frac{\gamma_0(i)}{1 - \sum_{j=1}^{m} \gamma_j(i)}. \tag{3.11}$$

It is worth noting the role of Assumption 1.b which guarantees the existence of the inverse $(1 - \sum_{j=1}^{m} \gamma_j(i))^{-1} \geqslant (1 - \Gamma)^{-1}$ for each action $i \in [k]$. The proof can be found in Appendix A.2.

Setting To demonstrate the importance of the noise in this setting, we consider the two clairvoyant policies defined above. We compare the optimal `Stochastic` policy (Equation 3.5) and the optimal policy for the `Deterministic` setting (Equation 3.11). The setting selected is challenging and made of $k = 2$ actions, i_1 and i_2, that are very close in terms of expected steady-state reward:

$$\boldsymbol{\gamma}(i_1) = (1,\ \rho,\ 0)^T \qquad\qquad \boldsymbol{\gamma}(i_2) = (1,\ 0,\ \rho - \epsilon)^T,$$

where $\rho = 0.5$, $\epsilon = 0.02$ and the noise is Gaussian with $\sigma \in \{0,\ 0.1,\ 0.5,\ 1.0,\ 2.0\}$.

Results Table 3.2 shows the performance of the two policies in terms of cumulative reward. First, with no noise (*i.e.*, $\sigma = 0$), the performances of the two policies are equivalent. However, when we consider a stochastic setting (*i.e.*, $\sigma > 0$), the `Stochastic` policy can exploit the beneficial effect of the noise in order to increase the average reward. Indeed, the optimal `Deterministic` policy retrieves almost the same reward for all the tested values of σ, while `Stochastic` policy increases its average reward as much as the system is noisy (since it can exploit it).

On the Knowledge of Parameter g

A fundamental parameter of AR−UCB is the value $g = \max_{i \in [k]} \gamma_0(i)$. In this part, we empirically show that any choice in the same order of magnitude as the actual value will let the algorithm achieve a sublinear regret, while severe underestimation prevents the algorithm from achieving a sublinear cumulative regret.

Setting We run multiple simulations varying the value of parameter $\overline{g}$. We chose $k = 7$, $m = 4$ and $\gamma_0(i) = 500$ for every action $i \in [k]$ (*i.e.*, $g = 500$). The autoregressive parameters $\gamma_j(i)$ have been sampled from a uniform probability distribution with support in $[0, 1/4 - \epsilon]$, where $\epsilon > 0$ is an arbitrarily small value. For this experiment, we test values $\overline{g} \in \{1, 10, 100, 500, 1000, 2500\}$.

Results In Figure 3.2, we report the cumulative regret of AR−UCB under different choices of $\overline{g}$. First, it is worth noting how choosing values of $\overline{g} \geqslant g$ always results in a sublinear cumulative regret, with a progressive increase as $\overline{g}$ gets larger. This is highlighted when comparing the scenario where $\overline{g} = 2500$ to the one where $\overline{g} \in \{500, 1000\}$. When $\overline{g}$ is underestimated, we empirically observe two facts. When $\overline{g}$ is in the same order of magnitude as the true value g (*e.g.*, $\overline{g} = 100$), we empirically observe a smaller sublinear cumulative regret. Instead, a severe underestimation of the parameter leads to a linear cumulative regret, as clearly visible for $\overline{g} \in \{1, 10\}$, although, in these settings, the cumulative regret is lower w.r.t. the other settings in the very first stages of the simulations (due to a more limited exploration).

On the Knowledge of the Autoregressive Order k

As discussed in Section 3.1.5, AR−UCB can also run under a misspecified parameter $\overline{k} \neq k$. We now empirically study the effect of misspecifying such a value.

Setting We consider a configuration with $n = 7$, $k = 10$, $\gamma_0(a) = 1$ and $\gamma_i(a)$ for $i \geqslant 1$ sampled from a uniform distribution having support in $[0, 10^{-2} \cdot 2i)$ for every action $a \in \mathcal{A}$. AR−UCB is run varying the parameter $\overline{k} \in \{1, 2, 4, 8, 10, 16\}$.

Results Figure 3.3 reports the average cumulative regret for the considered values of $\overline{k}$. On the one hand, an underestimation of parameter k (*i.e.*, $\overline{k} \in \{1, 2, 4\}$) results in an asymptotically linear cumulative regret. This effect is justified since AR−UCB is not able to learn the actual AR dynamics due to underfitting, *i.e.*, the considered models are too simple. On the other hand, AR−UCB achieves sublinear cumulative regret when $\overline{k} \geqslant k$ (*i.e.*, $\overline{k} \in \{10, 16\}$). In particular, when $\overline{k} > k$, the linear models use more parameters than required, resulting in slower learning. However, as the samples increase, the algorithm learns that the exceeding coefficients are not significant. A particular case is when $\overline{k}$ is close to k but strictly lower (*i.e.*, $\overline{k} = 8$). Here, the cumulative regret degenerates to linear, but if the coefficients $\gamma_j(a)$ for $j \in [\![\overline{k} + 1, k]\!]$ are not

very large, the performance of AR−UCB with misspecified $\overline{k}$ results, in practice, close to the one obtained with the true k.

3.1.7. Future Directions

In this section, we faced the online sequential decision-making problem where an autoregressive temporal structure between the observed rewards is present. First, we formally introduced the ARB setting and defined the notion of optimal policy, demonstrating that a myopic policy is optimal also to optimize the total reward, regardless of the target time horizon, and that the optimal policy is not constant over time and depends on the realizations of the reward. Then, we proposed an optimistic bandit algorithm, AR−UCB, to learn online the parameters of the underlying process for each action. We demonstrated that the presented algorithm enjoys sublinear regret, depending on the AR order k and on an index of the speed at which the system reaches a stable condition. Interestingly, the ARB setting can be casted as a special type of MDP. Nonetheless, we characterized the optimal policy in closed form and provided an algorithm capable of achieving sub-linear finite-time regret. Finally, we provided an experimental campaign to validate the proposed solution, and we analyzed the behavior of AR−UCB when key parameters are misspecified. Future directions should focus on fully understanding the complexity of learning in the ARB setting, deriving tight lower bounds, and matching algorithms.

3.2. Bridging Rested and Restless Bandits

Rested and Restless Bandits are two well-known bandit settings that are useful to model real-world sequential decision-making problems in which the expected reward of an arm evolves over time due to the actions we perform or due to the nature. In this section, we propose Graph-Triggered Bandits (GTBs), a unifying framework to generalize and extend rested and restless bandits. In this setting, the evolution of the arms' expected rewards is governed by a graph defined over the arms. An edge connecting a pair of arms (i, j) represents the fact that a pull of arm i triggers the evolution of arm j, and vice versa. Interestingly, rested and restless bandits are both special cases of our model for some suitable (degenerated) graph. As relevant case studies for this setting, we focus on two specific types of monotonic bandits: rising, where the expected reward of an arm grows as the number of triggers increases, and rotting, where the opposite behavior occurs. For these cases, we study the optimal policies. We provide suitable algorithms for all scenarios and discuss their theoretical guarantees, highlighting the complexity of the learning problem concerning instance-dependent terms that encode specific properties of the underlying graph structure.

This section presents Genalti et al. (2024c) and Genalti et al. (2024b), joint projects with Marco

Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni and Alberto Maria Metelli. Genalti et al. (2024c) is published at the *International Conference on Machine Learning (ICML)*, and Genalti et al. (2024b) is currently under review in a specialistic journal.

Introduction

In the basic stochastic Multi-Armed Bandit (MAB, Lattimore and Szepesvári, 2020) problem, at each round, the learner is asked to choose an action (a.k.a. arm) among a finite action set and, then, it observes a reward drawn from an unknown probability distribution. The simplicity of the MAB framework is both a strength and a limitation. On the one hand, the simple nature of the framework allows for the development of elegant and efficient algorithms that can be exactly characterized and studied from an information-theoretic perspective. On the other hand, the basic MAB model assumes a relatively simplistic environment that may not capture the complexities of real-world situations. As a result, traditional MAB approaches might not be sufficient for more intricate decision-making problems where additional factors come into play. To address these limitations, researchers have extended the MAB framework by incorporating additional structures and complexities to handle realistic scenarios. Examples of that are *linear* (Abbasi-Yadkori et al., 2011), *continuous-action spaces* (Kleinberg et al., 2008), and *kernelized* bandits (Chowdhury and Gopalan, 2017), which presents structure over the arms, *non-stationary* bandits (Gur et al., 2014), which allow us to consider evolving environments, *delayed* reward bandits (Pike-Burke et al., 2018), allowing us to consider delayed feedback. Over the different structures available in the literature, we focus on these two specific types of MAB structures, called *restless* and *rested* bandits (Tekin and Liu, 2012). In the former, the expected rewards evolve following the time (*i.e.*, as an effect of *nature*); in the latter, the expected reward of an arm evolves as a function of the pulls we perform on that specific arm.

In this section, we propose a unified framework to generalize restless and rested bandits. In particular, we define a novel space of MABs called Graph-Triggered Bandits (GTBs). A GTB is represented by a bandit complemented with a *graph* describing the interactions between arms. Specifically, an arm *triggers* the evolution of its own expected reward (as for rested bandit) and the evolution of the "connected" arms. Figure 3.4 shows an example of this scenario, where the nodes represent arms, and the edges represent interactions. Interestingly, rested and restless bandits are two vertices in the space of GTBs. In particular, restless bandits correspond to the case of a *fully-connected* graph, while rested ones correspond to the graph with the *self-loops only*.

This framework is driven by both *theoretical* and *practical* considerations. Theoretically, it offers a unified approach that generalizes *both rested and restless* bandits. Specifically, our goal is to

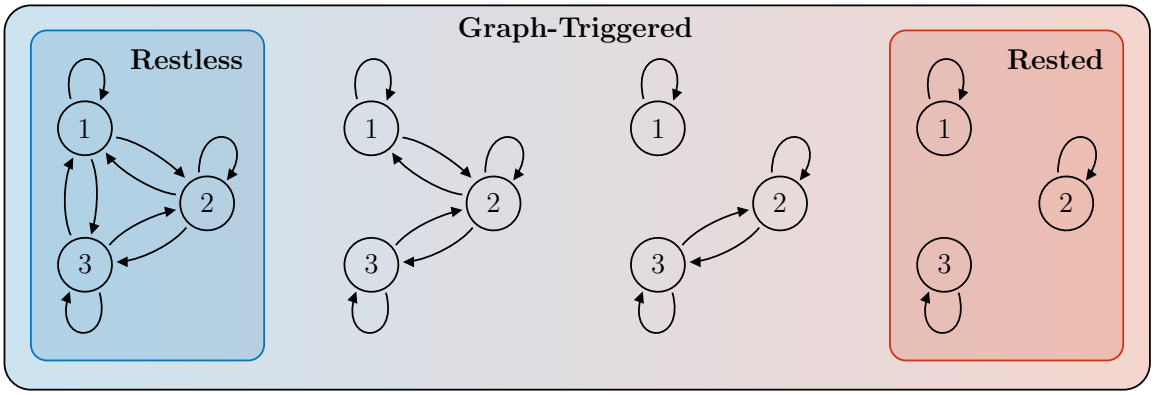


Figure 3.4: Examples of 3-armed GTBs.

establish a framework in which the well-known rested and restless bandits emerge as special cases, represented by appropriate (degenerated) graphs (see Figure 3.4). Practically, restless and rested bandits can model a wide range of real-world situations. For example, consider the scenario where we must choose which product to advertise (represented by our arms), with the reward being the number of sales for that product. On the one hand, with rested bandits, we can handle cases in which the products are all independent. On the other hand, with restless bandits, we can handle scenarios in which all the products interact. However, all the intermediate scenarios, *e.g.*, where advertising a product boosts its sales and also enhances the sales of the subset of complementary products, cannot be handled using restless/rested solutions. Indeed, this scenario is a rested problem with elements exhibiting restless behavior, and our generalization allows us to address such situations.

Contributions. In this section, we present Graph-Triggered Bandits (GTBs), a setting aiming to generalize and extend rested and restless bandits settings by introducing a graph structure to represent the interaction between the arms. We focus on the cases of rising and rotting bandits, as they represent interesting case studies allowing us to obtain no-regret algorithms. More in detail, the contributions are as follows.

- In Section 3.2.1, after having introduced the fundamental notions on the rested and restless bandits, we introduce the novel framework of GTBs and discuss the relevant quantities characterizing an instance, including a representation of the graph based on the connectivity matrix. Then, we present the learning problem and the performance index to evaluate algorithms in this setting.
- In Section 3.2.3, we study the *Rising GTBs* scenario. We discuss the optimal policy in this setting, by first providing a negative result, showing that computing the optimal policy is NP-hard for an arbitrary graph (Theorem 6). Then, we characterize the optimal policy for *block-diagonal* connectivity matrices, which can be computed in polynomial time (Theorem 7).

Subsequently, we discuss the deterministic scenario, and we propose two algorithms, the first, `DR-BD-UB`, for block-diagonal connectivity matrices and the second, `DR-G-UB`, for general graphs. We analyze their regret guarantees, highlighting the dependence on the graph structure (Theorems 8 and 9). Finally, we analyze the `R-▫-UCB` algorithm (Metelli et al., 2022), designed for rested and restless stochastic rising bandits that does not require the knowledge of the graph. We characterize its regret guarantees, focusing on the dependence on the characteristics of the underlying graph (Theorems 11 and 12).

- In Section 3.2.4, we study the *Rotting GTBs* scenario. As for the Rising GTBs case, we prove that computing the optimal policy is NP-hard for arbitrary graphs (Theorem 13). Then, we characterize the optimal policy for *block-diagonal* connectivity matrices, which admits a convenient closed-form solution (Theorem 14). Then, we focus on the special case of block-diagonal connectivity matrices, and we study how the `RAW-UCB` algorithm (Seznec et al., 2020) obtains strong regret guarantees with no knowledge of the graph (Theorem 15). Finally, we present a non-learnability result for *all* the Rotting GTBs problem under general matrices (Theorem 16).

The relevant literature is discussed in Section 3.2.2. The proofs of all the statements are provided in Appendices B.1 and B.2 for the Rising and Rotting GTBs, respectively.[7]

3.2.1. Graph-Triggered Bandits

In this section, we present the framework of Graph-Triggered Bandits (GTBs). We start in Section 3.2.1 by introducing the basic background notions on stochastic rested and restless bandits. Then, in Section 3.2.1, we formalize the GTBs setting. Finally, in Section 3.2.1, we formalize the learning problem for the GTBs setting.

Notions on Rested and Restless Bandits

We consider two specific types of MAB, namely *restless* and *rested* bandits (Tekin and Liu, 2012). In both cases, to each arm $i \in [k]$ corresponds a sequence of probability distributions $\boldsymbol{\nu} = (\nu_{i,n})_{i \in [k],\, n \in [T]}$, where the expected reward $\mu_i(n) = \mathbb{E}_{X \sim \nu_{i,n}}[X]$ evolves following an history-dependent quantity $n \in \mathbb{N}$. In the rested scenario, the expected reward of a generic arm i evolves according to the number of pulls of such an arm, i.e., $n \leftarrow N_{i,t}$. Conversely, in the restless case, the expected reward of a generic arm i evolves according to the current time t, i.e., $n \leftarrow t$. This means that, in rested bandits, the reward distribution of an arm evolves only when it is pulled, while in restless bandits, it evolves at each round, no matter the action performed. As

[7] For Rising GTBs, we report a short version of the proofs. The extended version is provided in (Genalti et al., 2024c).

customary in this field, we consider expected rewards $\mu_i(n)$ bounded in $[0,1]$, for every $i \in [k]$ and $n \in [T]$. Finally, we assume distributions to be *subgaussian* for every arm i and $n \in \mathbb{N}$, with their subgaussianity constants upper bounded by σ.

Setting

In rested and restless bandits there exists no structure among different arms. We now present a generalization of rested and restless bandits obtained by adding a structure allowing arms to interact. We consider arms as connected through an undirected graph, that can be either *known* or *unknown* to the agent.[8] If we pull an arm $i \in [k]$, we get its reward, and we *trigger* an evolution of the expected reward of the arm i and of all the arms connected to i. We do not get nor observe rewards from the connected arms (*i.e.*, bandit feedback). Such a graph can be represented by a symmetric Connectivity Matrix (CM) $\mathbf{G} \in \{0,1\}^{k \times k}$. If the matrix contains the value 1 in row i and column j, this implies that the pull of arm i determines the evolution of the expected reward of arm j. If the matrix contains a 0 in position (i,j), this implies that a pull of arm i does not cause an evolution of the expected reward of arm j. The pull of an arm i always implies the evolution of its own expected reward, formally $\mathbf{G}_{i,i} = 1, \ \forall i \in [k]$. For every round $t \in [T]$ and arm $i \in [k]$, we define the number $\widetilde{N}_{i,t}$ of *triggers* that it has undergone as follows:

$$\widetilde{N}_{i,t} = \sum_{\tau \in [t]} \mathbb{1}\{\mathbf{G}_{I_\tau,i} = 1\} = \mathbf{e}_i^\top \mathbf{G}^\top \mathbf{N}_t, \tag{3.12}$$

where $\mathbf{e}_i$ is a vector belonging to the canonical basis of $\mathbb{R}^k$ whose all components are all zero except for the i-th and $\mathbf{N}_t \coloneqq (N_{1,t}, \dots, N_{k,t})^\top$ is the vector containing the number of pulls of each arm up to round t. In GTBs, rewards are sampled from probability distributions whose average rewards vary with the number of triggers, i.e., $n \leftarrow \widetilde{N}_{i,t}$ and, consequently, the expected reward of an arm i evolves as $\mu_i(\widetilde{N}_{i,t})$. Furthermore, we define $t_{i,n} \coloneqq \sum_{l \in [T]} \mathbb{1}\{N_{i,l} \leqslant n\}$ as the round in which arm i has been pulled for the n-th time. With $\boldsymbol{t}_{i,t} \coloneqq (t_{i,n})_{n \leqslant N_{i,t}}$ we refer to the vector containing all the rounds in which the arm i has been pulled, up to time t. Moreover, we introduce $t^I_{i,n} \coloneqq \widetilde{N}_{i,t_{i,n}}$, namely the *internal time* of the n-th pull of arm i, which is the number of triggers of arm i at the time of the n-th pull. We also define, given the connectivity matrix of a graph $\mathbf{G}$, the notion of $\bar{k}_1 \coloneqq |\{i \in [k] : \deg(i) = 1\}|$ as the number of arms having degree of 1, where $\deg(i) \coloneqq \mathbf{1}_k^\top \mathbf{G} \mathbf{e}_i$ is the degree of a node, *i.e.*, the number of edges incident to the node. We now observe the relationship between rested and restless bandits and our setting.

Remark 5 (Inclusion of Rested and Restless bandits in GTBs)**.** *GTBs include both* rested *and* restless *bandits (Tekin and Liu, 2012). These two settings can be recovered by considering*

[8] All the results we present also hold for directed graphs.

$\mathbf{G} = \mathbf{I}_k$ *and* $\mathbf{G} = \mathbf{1}_{k\times k}$ *for rested and restless settings, respectively.*[9] *Indeed, a* restless *bandit can be seen as a particular instance of GTB where all arms are triggered at each round, making them change every round independently from which action has been chosen (*$\widetilde{N}_{i,t} = t$*, for every* $i \in [k]$*). Instead, in a* rested *bandit an arm changes its expected reward only when is directly chosen (*$\widetilde{N}_{i,t} = N_{i,t}$*, for every* $i \in [k]$*).*[10]

Block-Diagonal Connectivity Matrix. We now discuss a particular case of GTBs that is interesting from both the practical and analytical point of view. Until now, we considered $\mathbf{G} \in \{0,1\}^{k\times k}$ to be a general binary symmetric matrix. However, we now focus on the specific case in which $\mathbf{G}$ is a *block-diagonal* connectivity matrix, *i.e.*, a matrix in which the main-diagonal blocks are square matrices of all ones, and all off-diagonal blocks are zero matrices. Formally, let $\mathbb{B}_{\widetilde{k}} \subset \{0,1\}^{k\times k}$ be the set of block-diagonal connectivity matrices with exactly $\widetilde{k} \in [k]$ distinct blocks of 1s. We call the *GTBs with block-diagonal connectivity matrix* the set of instances where it holds that $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, for some $\widetilde{k} \leqslant k$. We identify with $\mathcal{C}_{\mathbf{G}} = \{C_{m,\mathbf{G}}\}_{m\in[\widetilde{k}]}$ the partition of $[k]$ corresponding to the diagonal blocks of $\mathbf{G}$. In graph theory, a block-diagonal connectivity matrix $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ corresponds to a cluster graph, *i.e.*, a graph formed from the disjoint union of complete graphs or *cliques* (Shamir et al., 2004). We call $\mathcal{C}_{\mathbf{G}}$ the set of cliques and we indicate with $\widetilde{N}_{C_m,t} := \sum_{i\in C_m} N_{i,t}$ the number of times an arm belonging to clique $C_m \in \mathcal{C}_{\mathbf{G}}$ has been pulled, namely the number of triggers of the clique C_m.

Connection to MDPs

The GTB interaction protocol can be seen as a special class of MDPs. We have a finite (thus large) state space $\mathcal{S}$. We define the state $s_t = (\widetilde{N}_{i,t})_{i\in[k]}$ by the means of the number of triggers of every action. It is easy to see that, when the graph is known, the learner can observe s_t and then decide which action I_t to play. The expected reward from choosing I_t in the state s_t is $f(s_t, I_t) = \mu_{I_t}(\widetilde{N}_{i,t})$. The transition is deterministic, and the state is updated by increasing the triggering counters. Finally, the initial state is deterministic and set to $s_1 = \mathbf{0}_k$. Suppose the graph is unknown to the learner. In that case, the current state cannot be observed and we have a *Partially Observable MDP (POMDP)*, a particular class of MDP that presents several additional difficulties. Moreover, it is worth noticing that this MDP is *non-communicating*, *i.e.*, once a state is left it is not possible anymore to reach it.

[9] We denote $\mathbf{I}_k$ the identity matrix of dimension k and $\mathbf{1}_{k\times k}$ the square matrix of dimension k whose entries are all equal to 1.

[10] This can be easily seen by looking at Equation (3.12) considering $\mathbf{G} = \mathbf{I}_k$ and observing that the vector $\mathbf{e}_i$ selects the i-th element of vector $\mathbf{N}_t$.

Learning Problem

We formally define the learning goal for GTB. This definition is coherent with the one provided in 2, but specifically suited for the notation of GTB. For a given instance $\boldsymbol{\nu}$ of a GTB, the performance of a policy π is measured by the means of *expected cumulative reward* throughout T rounds, formally:

$$J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \coloneqq \mathbb{E}\left[\sum_{t\in[T]} \mu_{I_t}(\widetilde{N}_{I_t,t})\right],$$

where the expectation is taken over the randomness of both the environment and the policy/algorithm. A policy is *optimal* for instance $\boldsymbol{\nu}$, a connectivity matrix $\mathbf{G}$, and time horizon T if it maximizes the expected cumulative reward, formally:

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_{\pi} J_{\boldsymbol{\nu},\mathbf{G},T}(\pi).$$

We denote by $J^*_{\boldsymbol{\nu},\mathbf{G},T} = J_{\boldsymbol{\nu},\mathbf{G},T}(\pi^*_{\boldsymbol{\nu},\mathbf{G},T})$ the expected cumulative reward attained by the optimal policy. We can now define the *expected policy regret* as:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) = J^*_{\boldsymbol{\nu},\mathbf{G},T} - J_{\boldsymbol{\nu},\mathbf{G},T}(\pi).$$

Therefore, our learning problem is to find a policy π minimizing the expected policy regret $R_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$. Since the optimal policy depends simultaneously on $\boldsymbol{\nu}$, $\mathbf{G}$, and T, from now on, we consider an instance of the GTB problem the triple $(\boldsymbol{\nu}, \mathbf{G}, T)$, instead of the reward distributions $\boldsymbol{\nu}$ only.

Remark 6 (On the Chosen Notion of Regret)**.** *In GTBs, we consider a notion of* policy *regret (Dekel et al., 2012a). Indeed, in this setting, diverging from the optimal sequence of actions influences not only instantaneous regret but also leads to a sub-optimal history, implying future regret even when returning to an optimal policy from there on. This notion of regret, which shares similarities with the one of reinforcement learning, is more challenging to optimize.*

3.2.2. Related Works

In this section, we discuss the relevant literature for the ARB setting. We divide this appendix into two parts. First, we discuss the relevant works concerning graph structures. Then, we discuss the literature related to restless and rested bandits, with particular attention to rotting and rising bandits.

Graph Relationships in Bandits. The graph-triggered bandits setting has been introduced in this work. Thus, no prior literature is available on this setting. However, we mention similar settings that appeared in the last years. Herlihy and Dickerson (2023) propose the networked restless bandit setting. Despite some similarities with our setting, *e.g.*, the presence of a graph among arms, their action space and learning objectives radically differ from ours, and thus the two settings are not comparable. In (Jhunjhunwala et al., 2018), a restless bandit setting is proposed in which the graph structure is not explicit in the formulation; however, the authors develop a graph representation of the policies in the deterministic scenario. Their algorithm builds and exploits a graph in an online fashion. Once again, we cannot properly compare this setting to ours, despite some sparse similarities. Finally, we mention bandits with graph feedback (Alon et al., 2015). Despite this setting being conceptually different from ours since arms do not interact, we report it here just because it features graph topology-dependent bounds. We remark that in this case, the graph has not to be intended as a structure for arms interactions but rather as a feedback structure for the learner, in GTBs the feedback is purely bandit.

Rested and Restless Bandits. Restless and rested bandits are a well-established research field. Starting from the seminal paper by Whittle (1988) on restless bandits, several approaches have been proposed over the years to deal with non-stationary bandits (Tekin and Liu, 2012; Raj and Kalyani, 2017). Then, specialization of these settings such as *rising* (Metelli et al., 2022; Mussi et al., 2024b) and *rotting* (Levine et al., 2017) has been introduced. Over the last years, several works tackled rotting bandits (Levine et al., 2017; Seznec et al., 2019). Remarkably, (Seznec et al., 2020) provide a single algorithm for dealing with both rested and restless rotting bandits but show that in the rotting setting, achieving sub-linear regret is not possible when there are both rested and restless arms in the same instance. We remark that for any two-armed rotting bandit where one arm is rested and the other is restless, we can construct an (asymmetric) matrix $\mathbf{G}$ such that the instance can be mapped to a graph-triggered rotting bandit instance. This highlights a crucial difference between rotting and rising bandits for what concerns graph-triggering. Recently, literature studied a broader class of restless bandits called *smooth* bandits, which generalizes both rotting and rising bandits (Manegueu et al., 2021; Jia et al., 2023).

3.2.3. Rising Graph-Triggered Bandits

Among the various types of restless and rested bandits available in the literature, in this section, we focus on *Rising Bandits* (Heidari et al., 2016; Metelli et al., 2022). We first introduce the assumption of the rising setting and some useful quantities. Then, we discuss the optimality in this setting (Section 3.2.3). Subsequently, we discuss the regret minimization problem for both the deterministic (Section 3.2.3) and stochastic (Section 6) scenarios.

Rising bandits are a specific class of MABs in which the expected reward of each arm evolves in a non-decreasing and concave manner. The following assumption formalizes such behavior.

Assumption 2 (Non-decreasing and Concave Payoffs)**.** *Let $\boldsymbol{\nu}$ be an instance of a rising bandit, then, defining $\gamma_i(n) := \mu_i(n+1) - \mu_i(n)$ for every $i \in [k]$ and $n \in [T]$, it holds:*

$$\textit{Non-decreasing:} \quad \gamma_i(n) \geqslant 0,$$
$$\textit{Concave:} \quad \gamma_i(n-1) \geqslant \gamma_i(n).$$

The two parts of this assumption allow us to provide theoretical guarantees in both the restless and rested settings. Such guarantees cannot be provided without the concavity assumption (see Theorem 4.2 of Metelli et al., 2022). We call *Rising GTBs*, the instances of GTBs in which the expected rewards fulfill Assumption 2.

Instance Characterization. Assumption 2 ensures sufficient structure on the problem to allow for algorithms with provably strong theoretical guarantees. In this scenario, given an instance $\boldsymbol{\nu}$, we define the *total increment* as:

$$\Upsilon_{\boldsymbol{\nu}}(M, q) := \sum_{t \in [M-1]} \max_{i \in [k]} \gamma_i(t)^q,$$

where $M \in \mathbb{N}$ and $q \in [0, 1]$. This quantity figures in the (instance-dependent) analysis of algorithms and characterizes the difficulty of learning in instance $\boldsymbol{\nu}$.

Optimality in Rising GTBs

In this part, we discuss the notion of *optimality* for our learning problem. We first characterize the complexity of finding the optimal policy followed by the clairvoyant when both the expected values and the matrix $\mathbf{G}$ are *known*.

Theorem 6 (Complexity of finding the Optimal Policy in Rising GTBs)**.** *Compu- ting the optimal policy in Rising GTBs with general matrices $\mathbf{G}$ is NP-Hard.*

This theorem follows from a reduction to the NP-Hard problem of determining if a large clique in a given graph exists (Karp, 1972). Intuitively, given a graph (V, E), we build an instance in which the cumulative reward is maximum only if the learner plays a sequence of arms that are associated with vertexes in a clique. Theorem 6 implies that the class of problems of Rising GTBs is computationally harder than all restless bandits and rested rising bandits, for which the optimal policy can be computed in polynomial time (Heidari et al., 2016). Moreover, the optimal policy does not admit a simple closed-form representation. Thus, in general, the optimal

policy cannot be reduced to a greedy one or to a fixed-arm policy. The result highlights how this definition of optimal policy is closer to the one of MDPs rather than the one of standard bandit settings.

We now show how, for the special case of Rising GTBs with block-diagonal connectivity matrices, the optimal policy can be efficiently computed and admits a closed-form solution.

Theorem 7 (Optimal Policy in Rising GTBs with Block-Diagonal CM)**.** *For any instance* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *of Rising GTBs with* $\mathbf{G} \in \mathbb{B}_{\tilde{k}}$*, the optimal policy* $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ *is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j \in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_j(t), \qquad \forall t \in [T],$$

where $C^*_{\boldsymbol{\nu},\mathbf{G},T}$ *is the "best" cumulative reward clique:*

$$C^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_{C \in \mathcal{C}_{\mathbf{G}}} \sum_{t \in [T]} \max_{j \in C} \mu_j(t).$$

This result characterizes the optimal policy when the graph linking the actions is only composed only by cliques. In particular, the clairvoyant would play a greedy policy but always inside the same predefined subset of arms composing a clique. Naturally, the chosen clique would be the one having the maximum cumulative reward at the end of the trial. We point out how this policy "combines" the optimal policies from both rising rested bandits (corresponding to always playing the arm with the highest *cumulative* reward), and the optimal policy from rising restless bandits (the *greedy* policy, corresponding to always playing the arm with the highest *instantaneous* reward).

Deterministic Rising GTBs

In this part, we propose two novel algorithms to learn in *deterministic* Rising GTBs, *i.e.*, all instances of Rising GTBs where $\sigma = 0$. More in detail, in Section 3.2.3, we discuss the block-diagonal CM case, while in Section 5, we discuss the general scenario. The deterministic scenario allows for a better understanding of the complex structure of this setting since it *ignores* the statistical learning problem.

We start by introducing a novel biased estimator which, for every arm $i \in [k]$, propagates its reward function to the current time t by estimating the first derivative using the last two observations:

$$\bar{\mu}_i(t) := \mu(t^I_{i,N_{i,t-1}}) + (t - t^I_{i,N_{i,t-1}}) \frac{\mu(t^I_{i,N_{i,t-1}}) - \mu(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}}. \tag{3.13}$$

Algorithm 5: `DR-BD-UB`.

Input : Connectivity matrix $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$

1 **for** $t \in [T]$ **do**
2 Compute $\bar{\mu}_i(t)$ as in Equation (3.13) , $\forall i \in [k]$
3 Select $I_t \in \arg\max_{i \in [k]} \bar{\mu}_i(t)$
4 Play I_t and observe $\mu_{I_t}(\widetilde{N}_{I_t,t})$
5 **end**

This estimator relies on the concept of *internal time*. Internal times are particularly useful since they can separate the bias in two components:

$$t - t^I_{i,N_{i,t-1}} = \underbrace{(t - t^I_{i,N_{i,t}})}_{\text{(A)}} + \underbrace{(t^I_{i,N_{i,t}} - t^I_{i,N_{i,t-1}})}_{\text{(B)}}.$$

As we will see in Section 3.2.3, this decomposition assumes a particular meaning in instances where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$, where (A) represents the rested component of the bias, since $t^I_{i,N_{i,t}} = \widetilde{N}_{C_m, t_{i,N_{i,t}}}$ making it equivalent to the bias of a rested bandit where cliques are the arms; and (B) represents the restless component of the bias, since from arm i perspective $t^I_{i,N_{i,t}} = \widetilde{N}_{i,t}$ can be interpreted as the current time inside the clique.

Algorithm for Deterministic Rising GTBs with Block-Diagonal CMs

We now introduce `Deterministic Rising Block-Diagonal Upper Bound` (`DR-BD-UB`), an optimistic anytime regret minimization algorithm for deterministic Rising GTBs with block-diagonal connectivity matrix, whose pseudocode is provided in Algorithm 5. The algorithm takes as input the connectivity matrix $\mathbf{G}$ and employs the estimator presented in Equation (3.13). Then, after having initialized the counters of the number of pulls, it starts the interaction with the environment. At each round $t \in [T]$, it estimates (line 2) the $\bar{\mu}_i(t)$ for every $i \in [k]$ as in Equation (3.13) and plays greedy according to it (line 4).[11]

The following result provides the regret bound of `DR-BD-UB`, highlighting the impact of the graph topology.

Theorem 8 (`DR-BD-UB` Regret in Det. Rising GTBs with Block-Diagonal CMs)**.** *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ and $\sigma = 0$. Then,* `DR-BD-UB` *suffers a*

[11] At the beginning, the algorithm is required to play every arm 2 times in a round-robin fashion in order to be able to compute $\bar{\mu}_i(t)$.

regret bounded by:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) \leqslant \widetilde{\mathcal{O}}\Bigg(\inf_{q\in[0,1]} \Bigg\{ \underbrace{T^q \sum_{C_m\in\mathcal{C}} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)}_{\text{(A) Rested Bias Contribution}} + \\ + \underbrace{\sum_{C_m\in\mathcal{C}} |C_m| \widetilde{N}_{C_m,T}^{\frac{q}{1+q}} \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}}_{\text{(B) Restless Bias Contribution}} \Bigg\} \Bigg).$$

In this theorem, we report the result as a function of the number of triggers $\widetilde{N}_{C_m,T}$ of the cliques in order to better discuss the properties of the graph. However, this dependence can be removed by simply observing $\widetilde{N}_{C_m,T} \leqslant T$. This choice allows us to have an interesting discussion on the nature of this result w.r.t. the graph structure. First of all, we observe that we can separate two contributions to the regret: one coming from the rested behavior (part (A) of the bound) determined by the need for identifying the best clique, and the other from the restless behavior needed for identifying the best arm inside the clique (part (B) of the bound). If we compare this result to the bounds in Theorems 4.4 and 5.2 of (Metelli et al., 2022), we can notice how the shapes of the two contributions correspond. We also remark that, in the two corner cases, *i.e.*, rested and restless bandits, the regret bound is actually smaller and corresponds exactly to the bounds presented in (Metelli et al., 2022), even though this is not immediately visible in Theorem 8 because of a mathematical artifact of the proof.[12] In the bound, the graph topology emerges by means of cliques' sizes, that act as multiplicative constants. The major consequence is that having fewer cliques leads, in general, to a better bound. As intuition suggests, the rested scenario can lead to a worst-case bound in the first component (which is, by the way, the one having the greater order in T), and this can be seen by a simple application of Jensen's Inequality, and by noticing that $\Upsilon_{\boldsymbol{\nu}}$ is a concave function:

$$\sum_{C_m\in\mathcal{C}} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right) \leqslant k \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{T}{k} \right\rceil, q \right).$$

We remark that in the two corner cases, one of the two contributions vanishes, even though it cannot be directly seen in Theorem 8. However, since the restless regret has a better order than the rested one, graphs with fewer cliques may lead, in general, to better bounds. Unfortunately, to precisely quantify this property, one would need to know the exact shape of $\Upsilon_{\boldsymbol{\nu}}$ and to solve a difficult optimization problem.

[12] More details can be found in Remark 13 (Appendix B.1).

Algorithm for Deterministic Rising GTBs with General Matrices

After having studied the scenario of block-diagonal connectivity matrices, we now consider the case in which $\mathbf{G}$ can be arbitrary. Before introducing the algorithm, we need to define the concept of *block sub-matrix*.

Definition 1 (Block Sub-matrix). *Let $\mathbf{G} \in \{0,1\}^{k\times k}$ be a general matrix, a block-diagonal matrix $\mathbf{G}^L \in \mathbb{B}_{\tilde{k}}$ is a* sub-matrix *of $\mathbf{G}$ if it satisfies:*

$$\mathbf{G}_{i,j} - \mathbf{G}^L_{i,j} \geqslant 0, \quad \forall i,j \in [k]. \tag{3.14}$$

Moreover, we say that $\bar{\mathbf{G}}^L \in \mathbb{B}_{\tilde{k}}$ is maximal *if it also satisfies:*

$$\bar{\mathbf{G}}^L \in \underset{\mathbf{G}^L \text{ satisfying Eq. (3.14)}}{\arg\min} |\mathcal{C}_{\mathbf{G}^L}| .$$

Informally, $\mathbf{G}^L \in \mathbb{B}_{\tilde{k}}$ is a sub-matrix of $\mathbf{G}$ if its graph can be obtained by only removing 1s from $\mathbf{G}$. Finally, a maximal sub-matrix has the least number of cliques. Note that such a maximal sub-matrix is, in general, not unique.

For this algorithm, we need to introduce a novel estimator, based on sub-matrices, whose definition recalls the one of Equation (3.13):

$$\bar{\mu}^L_i(t) := \mu(t^{I,L}_{i,N_{i,t-1}}) + (t - t^{I,L}_{i,N_{i,t-1}}) \frac{\mu(t^{I,L}_{i,N_{i,t-1}}) - \mu(t^{I,L}_{i,N_{i,t-1}-1})}{t^{I,L}_{i,N_{i,t-1}} - t^{I,L}_{i,N_{i,t-1}-1}}, \tag{3.15}$$

where $t^{I,L}_{i,l} := \mathbf{e}_i^\top (\bar{\mathbf{G}}^L)^\top \mathbf{N}_{t_{i,l}}$ is the internal time w.r.t. a maximal sub-matrix $\bar{\mathbf{G}}^L$ of the actual matrix $\mathbf{G}$. Given this new estimator, we can generalize Algorithm 5 to attain comparable performance even for a general connectivity matrix $\mathbf{G}$. We introduce a generalization of `DR-BD-UB` called `Deterministic Rising General Upper Bound` (`DR-G-UB`), whose pseudocode is provided in Algorithm 6. The algorithm takes as input a generic matrix $\mathbf{G}$ and computes $\bar{\mathbf{G}}^L$. Then, the algorithm interacts with the environment as before and uses the estimator defined in Equation (3.15). In other words, `DR-G-UB` pretends to be interacting with a bandit with a graph defined by $\bar{\mathbf{G}}^L$. The following result characterizes the performances of `DR-G-UB`.

Theorem 9 (`DR-G-UB` Regret in Det. Rising GTBs with General Matrices). *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \{0,1\}^{k\times k}$ and $\sigma = 0$. Then,* `DR-G-UB`

Algorithm 6: DR-G-UB.

Input : Connectivity matrix $\mathbf{G}$

Compute maximal sub-matrix $\bar{\mathbf{G}}^L$ from $\mathbf{G}$
for $t \in [T]$ **do**
  Compute $\bar{\mu}_i^L(t)$ as in Equation (3.15), $\forall i \in [k]$
  Select $I_t \in \arg\max_{i\in[k]} \bar{\mu}_i^L(t)$
  Play I_t and observe $\mu_{I_t}(\widetilde{N}_{I_t,t})$
end

suffers a regret bounded by:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-G-UB}) \leqslant \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]} \left\{ T^q \sum_{C_m^L \in \mathcal{C}_{\bar{\mathbf{G}}^L}} |C_m^L| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m^L,T}}{|C_m^L|} \right\rceil, q \right) + \sum_{C_m^L \in \mathcal{C}_{\bar{\mathbf{G}}^L}} |C_m^L| \widetilde{N}_{C_m^L,T}^{\frac{q}{1+q}} \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m^L,T}}{|C_m^L|} \right\rceil, q \right)^{\frac{1}{1+q}} \right\} \right),$$

where $\bar{\mathbf{G}}^L \in \mathbb{B}_{\widetilde{k}}$ *is a maximal sub-matrix of* $\mathbf{G}$.

This result provides a formal justification to the intuition that the performance of Algorithm 6 can be bounded with the upper bound attained in a less favorable scenario, *i.e.*, a block-diagonal instance that is "closer" to the worst-case instance of a rested bandit. The regret bound of DR-G-UB can be found by applying Theorem 8 using the matrix $\bar{\mathbf{G}}^L$.

Remark 7 (Computational Complexity)**.** *Note that, even if the optimal policy in this setting for a general* $\mathbf{G}$ *is NP-hard to be retrieved, with* DR-G-UB*, we achieve sublinear regret w.r.t. the optimal policy with a polynomial-time algorithm. This is possible thanks to the ability of* DR-G-UB *to identify a convenient matrix* $\bar{\mathbf{G}}^L$ *that is subsequently adopted as a proxy of the real environment in order to play in a computationally efficient manner.*

Stochastic Rising GTBs

In this part, we focus on the *stochastic* Rising GTBs scenario. We characterize the performances of R-▫-UCB (Metelli et al., 2022), designed for rising rested and restless bandits, in the Rising GTBs setting for both the block-diagonal CMs (Section 5) and the general case (Section 5). We show that such an algorithm achieves good performances for a general $\mathbf{G}$. In particular, we develop a new proof strategy for the regret upper bound that makes graph-dependent terms explicit. We aim at obtaining a computationally efficient algorithm enjoying *sublinear regret* guarantees. Surprisingly, our analysis shows that R-▫-UCB not only enjoys sublinear regret for any connectivity matrix $\mathbf{G}$, but also that the graph-dependent quantities actually interpolate the regret between the two corner cases. Moreover, we show that there is no need to solve any

Algorithm 7: R−□−UCB.

Input : Subgaussianity proxy σ, confidence levels $\{\delta_t\}_{t\in[T]}$, window size $\epsilon \in (0, 1/2)$.

for $t \in [T]$ **do**
  Compute $\widehat{\mu}_i^{h_{i,t}}(t)$ as in Equation (3.16), $\forall i \in [k]$
  Select $I_t \in \arg\max_{i\in[k]} \widehat{\mu}_i^{h_{i,t}}(t) + \beta_i^{h_{i,t}}(t, \delta_t)$
  Play I_t and observe $X_{I_t,t}$
end

additional NP-Hard problem before or during the algorithm's executions, letting R−□−UCB keep it affordable computational costs, as in the two corner settings. Furthermore, in this case, the algorithm is completely *unaware* of the graph structure.

The algorithm employs a biased estimator which, for every arm i, propagates its reward function to the current round t by estimating the first derivative over the last $2h$ samples:

$$\widehat{\mu}_i^h(t) := \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(X_{i,t_{i,l}} + (t-l) \frac{X_{i,t_{i,l}} - X_{i,t_{i,l-h}}}{h} \right), \tag{3.16}$$

where $h \in \mathbb{N}$ is the window size. We report the estimator's concentration rate, which is a function of the window size h. The proof of this result originally appeared in (Metelli et al., 2022). However, it can be extended to Rising GTBs (more details are provided in Appendix B.1.1).

Lemma 10 (Concentration of Estimator, adapted from Metelli et al. 2022)**.** *For every arm $i \in [k]$, every round $t \in [T]$, and window width $1 \leqslant h \leqslant \left\lfloor \frac{N_{i,t-1}}{2} \right\rfloor$, let:*

$$\beta_i^h(t, \delta) := \sigma(t - N_{i,t-1} + h - 1) \sqrt{\frac{10 \log \frac{1}{\delta}}{h^3}}.$$

Then, if the window size depends on the number of pulls only $h_{i,t} = h(N_{i,t-1})$ and if $\delta_t = t^{-\alpha}$ for some $\alpha > 2$, it holds for every round $t \in [T]$ that:

$$\mathbb{P}\left(\left| \widehat{\mu}_i^{h_{i,t}}(t) - \widetilde{\mu}_i^{h_{i,t}}(t) \right| > \beta_i^{h_{i,t}}(t, \delta_t) \right) \leqslant 2t^{1-\alpha}.$$

Algorithm. The algorithm, whose pseudocode is reported in Algorithm 7, takes as input the subgaussianity proxy σ, the sliding window size parameter ϵ, and a sequence of properly selected confidence levels δ_t, where $t \in [T]$. R−□−UCB relies on the previously defined biased estimator and uses its confidence interval to make decisions in an optimistic manner. R−□−UCB does not require the time horizon T as an input, making it an anytime algorithm. Moreover, the

algorithm exploits the sliding window mechanism to deal with the environment's uncertainty while controlling the confidence degree by means of $\{\delta_t\}_{t\in[T]}$. In particular, the window size employed by the algorithm is proportional to parameter $\epsilon \in (0, 1/2)$, in the form of $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$. As we show below, ϵ controls the bias-variance trade-off, where low values for ϵ result in less bias but higher variance, and vice versa.

Remark 8 (Computational Complexity)**.** *At each round,* R-▫-UCB *only needs to update the estimator and the related confidence bounds for every arm, which can be done in a time linear in the number of arms at every step. For an efficient update, we refer the reader to (Mussi et al., 2024b, Appendix C).*

Regret for Stochastic Rising GTBs with Block-Diagonal CMs

We now analyze the performances of R-▫-UCB in the block-diagonal CMs case.

Theorem 11 (R-▫-UCB Regret in Rising GTBs with Block-Diagonal CMs)**.** *Let* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *be an instance of Rising GTB, where* $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$*. Let* $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ *for* $\epsilon \in (0, 1/2)$ *and* $\delta_t = t^{-\alpha}$ *for* $\alpha > 2$*. Then,* R-▫-UCB *suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\text{R-▫-UCB}) \leqslant \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]} \left\{ \underbrace{(\sigma T)^{\frac{2}{3}}}_{\text{(A) Variance Contribution}} + \underbrace{\bar{k}_1 T^q \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{T}{\bar{k}_1} \right\rceil, q \right)}_{\text{(B) Rested Bias Contribution}} + \underbrace{T^{\frac{2q}{1+q}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| > 1} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{T}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}}_{\text{(C) Restless Bias Contribution}} \right\} \right),$$

where $\bar{k}_1$ *is the number of cliques in* $\mathbf{G}$ *containing only one action.*

Existence of a Bias-Variance Trade-off. In the regret upper bound, we can observe three distinct contributions. First, (A) represents the variance contribution, which is the regret suffered by the algorithm due to the stochastic nature of the environment. This contribution is due to the estimator's concentration properties and sets a minimum order of regret to $\widetilde{\mathcal{O}}((\sigma T)^{2/3})$. This term is independent of the total increment $\Upsilon_{\boldsymbol{\nu}}$ but, differently from the others, is the only contribution depending on σ. The contribution due to the estimator's bias is split into two distinct parts. The term (B) represents the rested contribution, which scales with the number of blocks containing only one arm. The term (C), instead, represents the restless contribution that scales with the number and the sizes of cliques. The bias contributions depend explicitly on the shape of average reward functions by total increment $\Upsilon_{\boldsymbol{\nu}}$. The only term common to variance and bias contributions is ϵ. Indeed, ϵ regulates such a trade-off between bias and variance, and this effect can be observed in the complete expression of the regret upper bound in Appendix B.1.

The variance contribution depends linearly on ϵ^{-1}; thus, a smaller window size implies a higher variance in the estimate. On the contrary, the bias tends to increase with ϵ: this is expected since a larger window means including older samples in the estimate.

Dependence on Graph Topology. In the regret upper bound of Theorem 11, the only contributions depending on graph topology are the bias ones (terms (B) and (C)). Indeed, the environment's randomness contribution has been decoupled from the estimation bias to get a tractable stochastic structure. We observe how the different behaviors of arms not connected with the others (size-1 cliques, corresponding to rested arms) and arms belonging to larger cliques. The regret scales as T^q in rested arms, but the dependence on the total increment $\Upsilon_{\boldsymbol{\nu}}$ is linear. Instead, for cliques with size greater than 1, regret scales as $T^{\frac{2q}{1+q}}$, which is greater than in rested contribution, but scales with $\Upsilon_{\boldsymbol{\nu}}$ to the power of $\frac{1}{1+q}$, that is indeed a better dependence. Moreover, each clique contributes differently, based on its size. Overall, the higher the size, the higher the contribution is, since the linear term is dominant w.r.t. the inverse term inside the total increment $\Upsilon_{\boldsymbol{\nu}}$. Another interesting dependence is the one on ϵ^{-1} for the restless contribution, which can be observed in the complete form of the bound in Appendix B.1. For connected arms, stochasticity and graph topology produce an interaction. Indeed, if one could design an estimator with strong concentration properties for connected arms, this would simplify the analysis of the restless contribution, eliminating the bad dependence on stochasticity. With such an estimator, we conjecture we could reduce the dependence up to $T^{\frac{q}{1+q}}$, matching the deterministic setting bound.[13]

Comparison with Known Results from Literature. Given that rested and restless rising bandits are special instances of Rising GTBs, we now comment on how the presented bound links to existing results when Algorithm 7 is run over one of those instances. We start from the rested scenario, *i.e.*, when $\mathbf{G} = \mathbf{I}_k$. Then, we would have $\bar{k}_1 = k$ and an empty summation in the restless bias contribution. The bound of Theorem 11 would thus assume the following form:

$$R_{\boldsymbol{\nu},\mathbf{I}_k,T}(\texttt{R-}\square\texttt{-UCB}) \leqslant \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{(\sigma T)^{\frac{2}{3}} + kT^q\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil\frac{T}{k}\right\rceil, q\right)\right\}\right).$$

The only other existing result for the rested rising bandits setting is the one of Theorem 4.4 of (Metelli et al., 2022), which is matched up to constants by ours. In the restless scenario, *i.e.*, when $\mathbf{G} = \mathbf{1}_{k\times k}$, we have a unique clique of size k, and $\bar{k}_1 = 0$. Thus, the bound we presented

[13] The lower bounds for rising rested and restless bandits are still an open problem.

in Theorem 11 becomes:

$$R_{\boldsymbol{\nu},\mathbf{1}_{k\times k},T}(\texttt{R-}\square\texttt{-UCB}) \leqslant \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{(\sigma T)^{\frac{2}{3}} + kT^{\frac{2q}{1+q}}\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil\frac{T}{k}\right\rceil, q\right)^{\frac{1}{1+q}}\right\}\right).$$

Once again, this result matches (up to constants) the result from Theorem 5.3 of (Metelli et al., 2022), the current state-of-the-art for the restless rising bandits problem. To conclude, we generalize the stochastic rising rested/restless bandit setting, with regret bounds that are tight w.r.t. the known results for the two corner scenarios.

Regret for Stochastic Rising GTBs with General Matrices

We are now ready to generalize the result of Theorem 11 to general matrices in $\mathbf{G} \in \{0,1\}^{k\times k}$. Before that, we have to first introduce the notion of *block super-matrix*.

Definition 2 (Block Super-matrix). *Let $\mathbf{G} \in \{0,1\}^{k\times k}$ be a general matrix, a block-diagonal matrix $\mathbf{G}^U \in \mathbb{B}_{\widetilde{k}}$ is a* super-matrix *of $\mathbf{G}$ if it satisfies:*

$$\mathbf{G}_{i,j} - \mathbf{G}^U_{i,j} \leqslant 0, \quad \forall i,j \in [k]. \tag{3.17}$$

Moreover, we say that $\bar{\mathbf{G}}^U \in \mathbb{B}_{\widetilde{k}}$ is minimal *if it also satisfies:*

$$\bar{\mathbf{G}}^U \in \underset{\mathbf{G}^U \text{ satisfying Eq. (3.17)}}{\arg\max} \left|\mathcal{C}_{\mathbf{G}^U}\right|.$$

This concept of minimal super-matrix plays an analogous role as the maximal sub-matrix in Theorem 9. We now have all the elements to present the upper bound on the regret for the stochastic Rising GTBs case and general matrices.

Theorem 12 (R-□-UCB Regret in Rising GTBs with General Matrices). *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \{0,1\}^{k\times k}$. Let $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$ and $\delta_t = t^{-\alpha}$ for $\alpha > 2$. Then,* R-□-UCB *suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-}\square\texttt{-UCB}) \leqslant \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{(\sigma T)^{\frac{2}{3}} + T^q\bar{k}_1\Upsilon_{\nu}\left(\frac{T}{\bar{k}_1}, q\right) + T^{\frac{2q}{1+q}}\sum_{C^U_m}|C^U_m|\Upsilon_{\nu}\left(\frac{T}{|C^U_m|}, q\right)^{\frac{1}{1+q}}\right\}\right),$$

where $\bar{\mathbf{G}}^U$ is the minimal super-matrix of $\mathbf{G}$.

This result is obtained by bounding $\widetilde{N}_{C^U_m,T} \leqslant T$ for every $C^U_m \in \mathcal{C}_{\bar{\mathbf{G}}^U}$ to remove any stochastic quantity from the regret, but a more precise bound can be provided by finding the worst-case allocation of the triggers among the cliques (as discussed for the similar result in Theorem 9).

However, this would require solving a challenging optimization problem that does not admit any closed-form solution. This result is similar to the one presented in Theorem 9, with the only difference being that the dependence on graph topology is linked to the minimal super-matrix. In principle, the result holds for any super-matrix of $\mathbf{G}$. Still, in the stochastic setting, the upper bound for the rested scenario is better than the one for the restless scenario. Hence, a block-diagonal CM with as many cliques as possible will, in most cases, lead to better bounds.

About the Knowledge of $\mathbf{G}$. In the stochastic scenario, we avoid extracting the super-matrix structure from the graph before executing the algorithm, as it always plays the same policy, regardless of the graph structure. Indeed, Algorithm 7 *does not require the knowledge on the graph*: the algorithm plays as if the true matrix is the identity one (*i.e.*, a rested instance). To justify this behavior in an intuitive way, have to look at Theorems 4.4 and 5.3 of (Metelli et al., 2022): in stochastic scenarios, the *rested* contribution to regret's upper bound has a better dependence on T w.r.t. the restless one. Moreover, our optimistic estimator computed by assuming a less connected graph will always be higher than the one computed from any more densely connected graph. Thus, by playing a purely rested policy, we are always sure to over-estimate the true reward (*i.e.*, optimism holds) and we are guaranteed that the rested contribution to the regret is maximized w.r.t. the restless contribution. The final form of the regret bound is obtained by including the minimal super-matrix as a pessimistic proxy of the effect of connected arms (informally, the minimal super-matrix represents the maximum possible contribution to the regret that is due to the arms connections). We point out that Algorithm 7 does not require the minimal super-matrix as an input, as it is needed only in the analysis. For this reason, one could reformulate the following result by removing the dependence on the minimal super-matrix and including a minimization over the set of all super-matrices. As a side effect, this dramatically reduces the computational burden w.r.t. the deterministic setting at the cost of a slightly higher regret bound.

Comparison with Deterministic Regret Bounds. In deterministic scenario (Theorems 8 and 9), the restless contributions are always of smaller order compared to the rested one, which is the contrary of what we observe in stochastic settings (Theorems 11 and 12). Due to this reason, in Algorithm 6, the regret bound scales with the maximal sub-matrix instead of the minimal super-matrix. In the deterministic setting, the maximal sub-matrix represents the maximum possible contribution to the regret that is due to the *absence* of arms connections. In principle, we could remove the necessity for graph knowledge also in the deterministic setting by simply playing as in a rested scenario (*i.e.*, run Algorithm 5 by setting $\mathbf{G} = \mathbf{I}_k$). This would be sensibly sub-optimal since any graph connection can be used to obtain a strictly better regret bound. This is not the case for the stochastic setting, where over-estimating the number of connections (*e.g.*,

by playing as in a restless scenario) may result in a non-optimistic estimator, compromising the theoretical soundness of our algorithms.

3.2.4. Rotting Graph-Triggered Bandits

Rotting bandits (Levine et al., 2017) are an important family of evolving rewards bandits where, contrary to what happens in rising bandits, the reward functions are not allowed to grow. In this section, we explore how the graph-triggering mechanism interacts with the non-increasing reward function assumption. We characterize the optimal policies and the challenges in finding them (Section 3.2.4). Then, we study the regret minimization problem for this setting in the presence of stochastic noise (Section 3.2.4).[14] Before that, we start by stating the main setting assumption and presenting the quantities characterizing this specific kind of bandits.

Assumption 3 (Non-increasing Payoffs)**.** *Let $\boldsymbol{\nu}$ be an instance of a rotting bandit, then, defining $\gamma_i(n) := \mu_i(n+1) - \mu_i(n)$ for every $i \in [k]$ and $n \in [T]$, it holds:*

$$\textit{Non-increasing:} \qquad \gamma_i(n) \leqslant 0.$$

This assumption allows for strong theoretical guarantees in both the restless and rested settings, as it has been shown in the literature (see, *e.g.*, Heidari et al., 2016; Levine et al., 2017; Seznec et al., 2019, 2020). Notably, for rotting bandits, we are not required to have a concavity/convexity assumption.

Instance Characterization. In this scenario, given an instance $\boldsymbol{\nu}$, we define the *total decrement* as:

$$V_{\boldsymbol{\nu}}(M) := \sum_{n \in [M-1]} \max_{i \in [k]} \gamma_i(n),$$

where $M \in \mathbb{N}$. Moreover, we define the *maximum per-round variation* as:

$$L := \max_{i \in [k]} \max_{n \in [T]} |\gamma_i(n)|,$$

with $\mu_i(-1) := \max_{i \in [k]} \mu_i(0)$. These quantities figure in the instance-dependent guarantees of algorithms operating in this setting and characterize the difficulty of learning for the instance $\boldsymbol{\nu}$. In particular, $V_{\boldsymbol{\nu}}(T)$ is required to properly bound the regret in restless rotting bandits (see, *e.g.*, Seznec et al. 2020), while L appears in the minimax regret bound of rested rotting bandits, as shown in the setting's lower bound of $\Omega(kL)$ by Heidari et al. (2016).

[14] For Rotting GTBs, we skip the deterministic case, as all the interesting results we want to show are visible also in the presence of noise.

Optimality in Rotting GTBs

Under the standard literature's assumptions, we are now ready to characterize our optimal policies. We first show, as for Rising GTBs, a negative result on the complexity of finding the optimal policy for a clairvoyant who knows all about our Rotting GTBs instance.

Theorem 13 (Complexity of finding the Optimal Policy in Rotting GTBs)**.** *Computing the optimal policy in Rotting GTBs with general matrices* $\mathbf{G}$ *is NP-Hard.*

The proof of this result follows a similar logic to the one of Theorem 6. Given this result, we now proceed by studying the block-diagonal connectivity matrices scenario, which composes an interesting class of Rotting GTBs. We now characterize the optimal policy in the block-diagonal connectivity scenario and the total cumulative reward that it obtains.

Theorem 14 (Optimal Policy in Rotting GTBs with Block-Diagonal CM)**.** *For any instance* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *of Rotting GTBs s.t.* $\mathbf{G} \in \mathbb{B}_{\tilde{k}}$*, the optimal policy* $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ *is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j\in[k]} \mu_j(\widetilde{N}^*_{j,t}), \qquad \forall t \in [T],$$

where $\widetilde{N}^*_{j,t}$ *is the number of times arm* j *has been triggered by the optimal policy up to time* t*. Moreover, we have:*

$$J^*_{\boldsymbol{\nu},\mathbf{G},T} = \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i\in C_m} \mu_i(n), \tag{3.18}$$

where $N^*_{C_m,T}$ *is the number of times the optimal policy pulls an action belonging to clique* C_m *before* T*, i.e.,* $N^*_{C_m,T} = \widetilde{N}^*_{i,T}$*, for every* $i \in C_m$*.*

Interestingly, the optimal policy is *greedy* in every rotting bandit with block-diagonal CM: this extends known results for rested and restless rotting bandits, where the optimal policy was already be proven to be greedy (Heidari et al., 2016; Levine et al., 2017). Equation (3.18) provides a closed form of the total reward obtained by the optimal policy, which will come in handy in the next part.

Stochastic Rotting GTBs

In this part, we discuss the regret minimization problem for the stochastic Rotting GTBs. We first study an algorithm, namely `RAW-UCB` (Seznec et al., 2020), which is able to achieve sublinear regret in the block-diagonal connectivity scenario (Section 3.2.4). Then, we show that, under the literature's standard assumptions, we cannot learn for general matrices (Section 5).

Algorithm 8: RAW−UCB (Seznec et al., 2020)

Input : Subgaussianity proxy σ, confidence levels $\{\delta_t\}_{t\in[T]}$.

1 **for** $t \in [T]$ **do**
2 Compute $\widehat{\mu}_i^h(t)$ as in Equation (3.19), $\forall i \in [k],\ h \in [N_{i,t}]$
3 Select $I_t \in \arg\max_{i\in[k]} \min_{h\leqslant N_{i,t}} \widehat{\mu}_i^h(t) + c(h, \delta_t)$
4 Play I_t and observe $X_{I_t,t}$
5 **end**

Algorithm for Stochastic Rotting GTBs with Block-Diagonal CMs

We now show that the RAW−UCB algorithm (Seznec et al., 2020), whose pseudocode is provided in Algorithm 8, provides sublinear regret guarantees in the Rotting GTB setting with block-diagonal connectivity matrices. RAW−UCB does not require any knowledge on $\mathbf{G}$ and allows for efficient computation.[15]

The behavior of RAW−UCB is characterized as follows. At each round $t \in [T]$, the algorithm computes a family of estimators for every action (line 2). In particular, for every action $i \in [k]$ and for every window size $h_{i,t} \in [N^{\pi}_{i,t-1}]$, it computes:

$$\widehat{\mu}_i^h(t) := \frac{1}{h}\sum_{s=1}^{t-1} \mathbb{1}_{\{I_t=i \ \wedge \ N_{i,s}>N_{i,t-1}-h\}} X_{i,s}. \tag{3.19}$$

Then, for every action, the chosen window size is the one minimizing the upper confidence bound $\widehat{\mu}_i^h(t) + c(h, \delta_t)$ where $c(h, \delta_t) := \sqrt{2\sigma^2 \log(2\delta_t^{-1})/h}$ (line 3).[16] Proving the algorithm's guarantees in the rested and restless setting requires characterizing how it concentrates, as has been done for the base case in Lemma 2 of (Seznec et al., 2020). We extend this result to the Rotting GTBs setting, devising a concentration bound involving the number of triggers of an action. the result can be found in Lemma 44 (Appendix B.2). This result will play a key role in the regret analysis of RAW−UCB in the Rotting GTBs setting. We are now ready to state the regret upper bound of RAW−UCB in the Rotting GTBs for block-diagonal connectivity matrices.

Theorem 15 (RAW−UCB Regret in Rotting GTBs with Block-Diagonal CM)**.** *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of the Rotting GTBs, where $\mathbf{G} \in \mathbf{B}_{\tilde{k}}$. Let $\delta_t = t^{-\alpha}$ for $\alpha \geqslant 5$. Then,*

[15] More details on the computationally efficient version of RAW−UCB, namely EFF−RAW−UCB, can be found in (Seznec et al., 2020).

[16] At the beginning of the execution, we need a round-robin pull of the arms to initialize the estimators.

RAW-UCB *suffers an expected regret bounded as:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB}) \leqslant \widetilde{\mathcal{O}}\Bigg(\underbrace{k\left(\sigma\sqrt{\log T} + V_{\boldsymbol{\nu}}(T)\right)}_{\text{(A) Variance Contribution}} + \underbrace{L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + kL + \sigma\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}\left(\sqrt{\frac{|C_m|}{k}T}\right)}_{\text{(B) Rested Contribution}} + \underbrace{(\alpha\sigma)^{\frac{2}{3}}\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}\left(V_T^{\pi}\frac{|C_m|}{k}T^2\right)^{\frac{1}{3}}}_{\text{(C) Restless Contribution}}\Bigg).$$

Dependence on Graph Topology. In Theorem 15, we observe the same phenomenon of Theorem 11. Indeed, we have three components: (A) representing the fixed regret contribution that comes from the noise, (B) representing the contribution to the regret coming from the rested nature of the problem (i.e., the sub-optimality accrued by choosing an action in the wrong clique), and (C) representing the contribution coming from the restless nature of the problem, instead (i.e., the sub-optimality accrued by not choosing the best action in a clique). The separation between the latter two components becomes clear in the proof of the result:

$$\begin{aligned}\mathbb{E}[R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB})] &= \sum_{t=1}^{T}\mu_{i_t^*}(\widetilde{N}^*_{i_t^*,t}) - \mu_{I_t}(\widetilde{N}^{\pi}_{I_t,t}) \pm \max_{i\in C_{I_t}}\mu_i(\widetilde{N}^{\pi}_{I_t,t}) \\ &= \underbrace{\sum_{t=1}^{T}(\mu_{i_t^*}(\widetilde{N}^*_{i_t^*,t}) - \max_{i\in C_{I_t}}\mu_i(\widetilde{N}^{\pi}_{I_t,t}))}_{\leqslant \text{(B)} + k\sigma\sqrt{\log T}} + \underbrace{\sum_{t=1}^{T}(\max_{i\in C_{I_t}}\mu_i(\widetilde{N}^{\pi}_{I_t,t}) - \mu_{I_t}(\widetilde{N}^{\pi}_{I_t,t}))}_{\leqslant \text{(C)} + kV_{\boldsymbol{\nu}}(T)}.\end{aligned}$$

In rotting bandits, there is a clear hierarchy between the difficulties of statistical learning in the rested and the restless settings. Rested rotting bandits are easier than their restless counterparts, and this is reflected also in our bound: when the number of cliques is higher, and the GTB is closer to a rested instance, the regret bound is lower since the weight of (A) is higher.

Comparison with Known Results from Literature. RAW-UCB has already been proven to be nearly optimal in both the rested and the restless scenario (Seznec et al., 2020). However, as an artifact of the analysis, we cannot retrieve the exact same bounds by plugging $\mathbf{G} = \mathbf{I}_k$, or $\mathbf{G} = \mathbf{1}_{k\times k}$ in our expression. A similar consideration to the one of Remark 13 can also be done for rotting bandits. We conjecture that RAW-UCB is actually nearly-optimal also in the intermediate Rotting GTB instances with block-diagonal connectivity matrices, and this claim is supported by the fact that there is no room for improvement in neither of the two contributions in the corner cases. We also conjecture that the dependence on $L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2$ may be an artifact of the analysis, being the output of a delicate pigeonhole principle argument used to

prove Theorem 15 (see Appendix B.2). We leave the task of finding a graph-dependent lower bound for this setting as a fascinating open problem.

Non-learnability for Stochastic Rotting GTBs with General Matrices

We now move to the scenario of stochastic Rotting GTBs for general connectivity matrices, and we present an impossibility result for this case.

Theorem 16 (Regret Lower Bound for Rotting GTBs with General Matrices)**.** *For every* $\mathbf{G} \in \{0,1\}^{k\times k}$ *that is not block-diagonal, there exists an instance of Rotting GTB* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *s.t., for every policy* π*, it holds:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \geqslant \frac{T}{12}.$$

This negative result poses a several limitation to what can be obtained from Rotting GTBs in the general scenario since, if we do not consider additional assumptions on the reward functions, no algorithm can obtain sublinear regret.

3.2.5. Future Directions

In this paper, we proposed *graph-triggered bandits* (GTBs), a generalization of the rested and restless bandit settings, where the expected rewards of the different arms evolve by means of a graph. We focused and compared two special families of bandits, namely *rising* and *rotting* bandits, where the expected rewards of an arm evolve in a monotonic fashion with the number of *triggers* the specific arm received. We showed that computing the optimal policy without additional assumptions on the matrix is NP-Hard in both the rising and rotting scenarios. Then, for both these classes, we showed how, instead, for block-diagonal connectivity matrices, we can find the optimal policy in polynomial time and have a convenient closed-form expression. From the algorithmic perspective, we showed how it is possible to achieve sublinear regret for both of these special instances of MABs with block-diagonal connectivity matrices. On the other hand, for general matrices, we have an interesting distinction. Indeed, for Rising GTBs we were able to achieve sublinear regret, while for Rotting GTBs (in the standard scenario, without additional assumption on the behavior of expected rewards, *e.g.*, on second-order derivatives), we proved that we cannot learn. this section aspires to be a first step in the study of GTBs and should be integrated by studying the statistical complexity of learning through lower bounds, and by considering more general models, *e.g.*, smoothly evolving bandits.

3.3. Tightening Regret Lower and Upper Bounds in Restless Rising Bandits

In this section, we delve deeper into the restless rising bandit problem, which has already been introduced in the previous section in the more general setting of graph-triggered bandits.

Restless Multi-Armed Bandits (MABs) are a general framework designed to handle real-world decision-making problems where the expected rewards evolve over time, such as in recommender systems and dynamic pricing. In this work, we investigate from a theoretical standpoint two well-known structured subclasses of restless MABs: the *rising* and the *rising concave* settings, where the expected reward of each arm evolves following an unknown *non-decreasing* and a *non-decreasing concave* function, respectively. By providing a novel methodology of independent interest for general restless bandits, we establish new lower bounds on the expected cumulative regret for both settings. In the rising case, we prove a lower bound of order $\Omega(T^{2/3})$, matching known upper bounds for restless bandits; whereas, in the rising concave case, we derive a lower bound of order $\Omega(T^{3/5})$, proving for the first time that this setting is provably more challenging than stationary MABs. Then, we introduce `AutoRegressive Upper Confidence Bound` (`AR-UCB`), a new regret minimization algorithm designed for the rising concave MABs. By devising a novel proof technique, we show that the expected cumulative regret of `AR-UCB` is in the order of $\widetilde{\mathcal{O}}(T^{7/11})$. These results collectively make a step towards closing the gap in rising concave MABs, positioning them between stationary and general restless bandit settings in terms of statistical complexity.

This section presents Migali et al. (2025), a joint project with Cristiano Migali, Marco Mussi and Alberto Maria Metelli, published at the *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

3.3.1. Introduction

The standard MAB setting considers stationary reward distributions. However, in many real-world decision-making problems, the expected rewards of available actions can vary over time due to changes in the surrounding environment, such as shifting in consumer preferences for online marketplaces (Wu et al., 2018) or evolving health status of patients in treatment selection during clinical trials (Aziz et al., 2021). To address such dynamics, the *restless* MABs framework (Tekin and Liu, 2012) has been introduced. This model generalizes the classical MAB setting by explicitly incorporating the *non-stationarity* of the arms.[17]

[17]With a slight abuse of terminology, we will use the words *non-stationary* and *restless* interchangeably.

		Holds for			Result
		Restless	Restless Rising	Restless Rising Concave	
Lower Bounds	Lattimore and Szepesvári 2020 (Thm. 15.2)	✓	✓	✓	$\Omega(T^{1/2})$
	Besbes et al. 2014 (Thm. 1)	✓	✗	✗	$\Omega(T^{2/3})$
	This thesis (Thm. 18)	✓	✓	✗	$\Omega(T^{1/2}) \to \Omega(T^{2/3})$
	This thesis (Thm. 19)	✓	✓	✓	$\Omega(T^{1/2}) \to \Omega(T^{3/5})$
Upper Bounds	Besbes et al. 2014 (Thm. 2)	✓	✓	✓	$\mathcal{O}(T^{2/3})$
	Metelli et al. 2022 (Thm. 5.3)	✗	✗	✓	$\widetilde{\mathcal{O}}(T^{2/3})$
	This thesis (Thm. 23)	✗	✗	✓	$\mathcal{O}(T^{2/3}) \to \widetilde{\mathcal{O}}(T^{7/11})$

Table 3.3: Existing and new bounds for the *restless*, *restless rising* and *restless rising concave* settings.
The arrow → points from the previous best result to the improved one presented in this paper.

Non-stationarity in bandit problems has been addressed through a variety of models and methods, such as restless bandits with *abrupt changes* in the reward distribution (*e.g.*, Garivier and Moulines, 2011), *smoothly* evolving expected rewards (*e.g.*, Trovò et al., 2020), and settings where the *total variation* of expected rewards is bounded over time (*e.g.*, Besbes et al., 2014). These frameworks allow the expected rewards to fluctuate in complex ways, such as increasing and then decreasing, without constraints on their direction of change. In contrast, there are important classes of bandit models that enforce *monotonicity* on the expected rewards. These include *rising bandits* (Heidari et al., 2016; Metelli et al., 2022), where expected rewards are non-decreasing, and *rotting bandits* (Levine et al., 2017; Seznec et al., 2019, 2020), where they are non-increasing. Such models are well-suited for capturing structured real-world dynamics, including online model selection (Metelli et al., 2022), hyperparameter optimization (Mussi et al., 2024a), and recommendation systems (Levine et al., 2017).

Motivation. In this section, we focus on the restless *rising* bandits and restless *rising concave* bandits and we aim to characterize them from a theoretical standpoint since several fundamental questions remain unresolved. In the general restless bandit setting, where the expected rewards may vary over time with bounded variation over T rounds, the minimax regret is known to be lower bounded by $\Omega(T^{2/3})$ (Besbes et al., 2014).[18] However, no regret lower bound has been derived for the specific class of non-decreasing (rising) or non-decreasing concave (rising concave) restless bandits yet, making the classical lower bound for stationary bandits, $\Omega(T^{1/2})$ (Lattimore and Szepesvári, 2020, Thm. 15.2), the best available reference, and leaving the following question

[18]We use $\Omega(\cdot)$ and $\mathcal{O}(\cdot)$ to highlight the dependence on T in the lower and upper bounds, respectively, omitting constant factors. For upper bounds, we also use $\widetilde{\mathcal{O}}(\cdot)$ to suppress logarithmic dependencies on T too.

open.

Research Question 1 *Is it possible to conceive regret **lower bounds** for restless rising and restless rising concave bandits that are* strictly larger *than the $\Omega(T^{1/2})$ bound for stationary bandits?*

The currently available algorithms for restless rising bandits are those designed for general restless bandits with bounded variation, which achieve a regret upper bound of order $\mathcal{O}(T^{2/3})$ (Besbes et al., 2014). When incorporating concavity, more specific algorithms have been proposed (Metelli et al., 2022), but unfortunately, they fail to improve the regret order. This generates the following question.

Research Question 2 *Is it possible to devise algorithms for restless rising and rising concave bandits whose regret **upper bounds** are* strictly smaller *than the $\mathcal{O}(T^{2/3})$ bound for general restless bandits?*

Original Contribution. In this paper, we aim to provide an answer to the research questions presented above, making a step towards the complete statistical characterization of restless rising and restless rising concave bandits. The contribution is summarized as follows:

- In Section 3.3.4, we provide a *general recipe* for deriving regret lower bounds for restless bandits, which generalizes the construction of Besbes et al. (2014) and is of potential independent interest (Lemma 17). We then *specialize* this construction to the cases of rising and rising concave bandits. First, we derive a lower bound of order $\Omega(T^{2/3})$ for rising bandits, showing that this setting shares the same statistical complexity as general restless bandits (Theorem 18) and answering negatively to Question Research Question 2 for rising bandits. Second, for restless rising concave bandits, we show that the regret is at least of order $\Omega(T^{3/5})$, showing that this setting is more challenging than stationary MABs (Theorem 19). These results provide a positive answer to Question Research Question 1 for both settings.
- In Section 3.3.5, we present `Rising Concave Budgeted Exploration` (`RC-BE`(α)), a novel regret minimization algorithm for restless rising concave MABs, which extends `Budgeted Exploration` (Jia et al., 2023). By devising a novel analysis, we provide an upper bound on its regret of order $\widetilde{\mathcal{O}}(T^{7/11})$ (Theorem 23) with no requested knowledge of the learning horizon or of the total variation. This result improves upon the current best upper bound of order $\mathcal{O}(T^{2/3})$ and provides a positive answer to **Question 2** for rising concave bandits.

Numerical simulations are provided in Section 3.3.6. Related works are discussed in Appendix 3.3.3. Omitted proofs are provided in Appendices C.1 and C.2 for lower and upper bounds, respectively. A summary of known and new results presented in this paper is provided in Table 3.3.

3.3.2. Problem Formulation

We recall the definitions provided for graph-triggered rising bandits provided in Section 3.2.3, and note that restless bandits represent a corner case for graph-triggered bandits (by letting $\widetilde{N}_{i,t} = t$ for every $i \in [k]$ and $t \in [T]$). We provide some additional notation that will come handy in the next sections. We denote the random table with all possible rewards as $\boldsymbol{X} = (X_{i,t})_{i\in[k],t\in\mathbb{N}_{\geqslant 1}}$. For every arm $i \in [k]$, we define its expected reward $\mu_i : \mathbb{N}_{\geqslant 1} \to \mathbb{R}$ as the expectation of the reward obtained by pulling such arm, *i.e.*, $\mu_i(t) = \mathbb{E}_{X\sim\nu_i(t)}[X]$ and denote the vector of expected reward functions as $\boldsymbol{\mu} = (\mu_i)_{i\in[k]}$. We assume that the expected rewards are bounded in $[0, 1]$, and that the realizations are σ-subgaussian.

Rising Bandits. In this section, we will distinguish between rising bandits and rising *concave* bandits. Thus, for the sake of clarity, we will re-state the two assumptions separately.

Assumption 4 (Non-Decreasing expected reward)**.** *Let $\boldsymbol{\nu}$ be a restless MAB. For every arm $i \in [k]$ and round $t \in \mathbb{N}_{\geqslant 1}$, the function $\mu_i(t)$ is non-decreasing in t. In particular, defining the increments:*

$$\gamma_i(t) \coloneqq \mu_i(t+1) - \mu_i(t) \geqslant 0.$$

Assumption 5 (Concave expected reward)**.** *Let $\boldsymbol{\nu}$ be a restless MAB. For every arm $i \in [k]$ and round $t \in \mathbb{N}_{\geqslant 1}$, the function $\mu_i(t)$ is concave in t, i.e.:*

$$\gamma_i(t+1) - \gamma_i(t) \leqslant 0.$$

Formally, we call *restless rising* a restless MAB in which Assumption 4 holds, and *restless rising concave* a restless MAB in which both Assumptions 4 and 5 hold. From now on, we omit the adjective *restless* for the sake of conciseness.

Connection to MDPs

The restless bandit problem can be casted to an MDP in the same way as we did for graph-triggered bandits. The state space $\mathcal{S} = [T]$ is now simpler, as only the time governes the system's evolution. Of course, this imply a deterministic transition from state $s_t = t$ to state $s_{t+1} = t + 1$. The MDP is always fully-observable and once action I_t is selected at time t, the expected reward is provided by $f(s_t, I_t) = \mu_{I_t}(t)$. The initial state is trivially $s_1 = 1$. As in GTB, the MDP is however non-communicating, indeed once a state is left there is not way to get back to it.

Learning Problem

As customary in MABs, we define the performance of a policy π in a restless MAB $\boldsymbol{\nu}$ as the *expected cumulative reward* collected over the T rounds, formally:

$$J_{\boldsymbol{\nu}}(\pi, T) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\nu}} \left[\sum_{t=1}^{T} \mu_{I_t}(t) \right].$$

A policy $\pi^*_{\boldsymbol{\nu}}$ is *optimal* if it maximizes the expected cumulative reward: $\pi^*_{\boldsymbol{\nu}} \in \arg\max_{\pi} \{J_{\boldsymbol{\nu}}(\pi, T)\}$. In restless MABs, the optimal policy does not explicitly depend on T and consists of pulling in each round the arm with the highest expected reward: $\pi^*_{\boldsymbol{\nu}}(t) \in \arg\max_{i \in [k]} \mu_i(t)$ for every $t \in \mathbb{N}_{\geqslant 1}$. Denoting with $J^*_{\boldsymbol{\nu}}(T) \coloneqq J_{\boldsymbol{\nu}}(\pi^*_{\boldsymbol{\nu}}, T)$ the expected cumulative reward of an optimal policy, the suboptimal policies π are evaluated via the *expected cumulative regret*:

$$R_{\boldsymbol{\nu}}(\pi, T) \coloneqq J^*_{\boldsymbol{\nu}}(T) - J_{\boldsymbol{\nu}}(\pi, T). \tag{3.20}$$

Instances Characterization. To characterize an instance $\boldsymbol{\nu}$, we introduce the following quantity, namely the *cumulative increment*, defined for every $t_1, t_2 \in \mathbb{N}_{\geqslant 1}$ with $t_1 \leqslant t_2$ as:

$$\Upsilon_{\boldsymbol{\nu}}(t_1, t_2) \coloneqq \sum_{l=t_1}^{t_2 - 1} \max_{i \in [k]} \gamma_i(l).$$

The cumulative increment extends to an arbitrary interval with t_1 and t_2 as extremes the analogous notion $\Upsilon_{\boldsymbol{\mu}}(T, q)$ employed in (Metelli et al., 2022), restricting to $q = 1$. It is immediate to show that $\Upsilon_{\boldsymbol{\nu}}(t_1, t_2) \in (0, k]$ since $\Upsilon_{\boldsymbol{\nu}}(t_1, t_2) \leqslant \sum_{l=t_1}^{t_2-1} \sum_{i \in [k]} \gamma_i(l) \leqslant \sum_{i \in [k]} 1 = k$. Analogously to what is done in (Besbes et al., 2014), we consider the class of instances whose cumulative increment over the learning horizon T is bounded by a *variation budget* $V_T \in (0, k]$, which we assume known, formally $\Upsilon_{\boldsymbol{\nu}}(1, T) \leqslant V_T$. Then, we call $\mathcal{E}_{\mathrm{r}}(T, V_T)$ and $\mathcal{E}_{\mathrm{c}}(T, V_T)$ the set of rising MABs and rising concave MABs instances whose $\Upsilon_{\boldsymbol{\nu}}(1, T)$ satisfies the inequality above, respectively.

3.3.3. Related Works

Restless Bandits. In the original *restless* MAB setting, introduced by Tekin and Liu (2012), the evolution of the expected reward of each arm was described by a Markov chain. Several algorithms have been proposed to deal with this new framework, *e.g.*, `Restless-UCB` (Wang et al., 2020), which relies on the optimistic estimation of the transition kernel of the underlying chain. Over time, the term restless acquired a broader meaning, encompassing all bandits in which the expected reward changes as time passes. Such arbitrary evolution can be described

by a function that maps each round to the expected reward of a given arm. This is the type of restless bandit we target in this work. There are two families of methods to tackle restless MABs: *passive* (*e.g.*, Garivier and Moulines, 2011; Besbes et al., 2014; Auer et al., 2019; Trovò et al., 2020) and *active* (*e.g.*, Liu et al., 2018; Besson et al., 2022; Cao et al., 2019). Passive methods base their estimates on the recent feedback, forgetting obsolete observations. Active methods try to detect the changes in arms' expected rewards and use only the observations gathered after the last change. Among the most common passive approaches we find methods based on discounted rewards, *e.g.*, `D-UCB` (Garivier and Moulines, 2011), or adaptive sliding window, *e.g.*, `SW-UCB` (Garivier and Moulines, 2011). Both algorithms suffer a $\widetilde{\mathcal{O}}(T^{1/2})$ regret in the setting in which expected rewards change abruptly a fixed number of times over the time horizon, and such number is known. Auer et al. (2019) obtained a similar result in the same setting, without knowing the number of changes, by resorting on the doubling trick (Besson and Kaufmann, 2018). Another common setting is the one that allows the expected rewards to evolve arbitrarily, with the only constraint that the maximum absolute difference between the expected rewards of an arm in one round and the next, summed over the time horizon, is smaller than or equal to a variation budget V_T (Besbes et al., 2014). The `Rexp3` algorithm (Besbes et al., 2014), a modification of the `Exp3` (Auer et al., 2002b) policy, originally designed for adversarial MABs, shows a regret bound of $\mathcal{O}(T^{2/3})$ under the knowledge of the variation budget V_T. The need for the knowledge of such quantity has been removed by Chen et al. (2019) by means of the doubling trick. In (Trovò et al., 2020), an approach which combines a Thompson-Sampling-like algorithm with a sliding window, shows theoretical guarantees in both the abruptly and smoothly changing settings.

Rising Bandits. *Rising concave* MABs have been introduced in the deterministic setting by Heidari et al. (2016) and Li et al. (2020), where the rewards observed by the agent in each round are not affected by noise. In their formulation of the problem, the rewards of an arm are non-decreasing in the number of times such an arm has been pulled and satisfy the *decreasing marginal return* assumption, *i.e.*, the increment in the reward observed between one pull and the next of the same arm is non-increasing in the number of pulls. The online algorithm designed by Heidari et al. (2016) to minimize the regret relies on an optimistic estimate of the cumulative reward that can be obtained by pulling a given arm. Indeed, in this setting, Heidari et al. (2016) show that the optimal policy consists of repeatedly pulling the arm with the highest cumulative reward over the horizon. Li et al. (2020) use the rising concave MAB framework to model the problem of parameter optimization in machine learning and design an algorithm based on iterative elimination of unpromising arms that has good empirical performance. Cella et al. (2021) consider a setting in which the reward is increasing in expectation and the observations are affected by noise. However, in their framework, the expected rewards are constrained to

follow a specific parametric form known to the agent. The authors analyze the setting under both the regret minimization and best arm identification frameworks. Anyway, the given parametric form makes this setting not applicable to an arbitrary expected reward evolution that satisfies the non-decreasing assumption. Recently, a surge of approaches has been designed for addressing other learning problems in stochastic rising concave MABs, including regret minimization (*e.g.*, Metelli et al., 2022) and best arm identification (*e.g.*, Takemori et al., 2024; Mussi et al., 2024a). Finally, Genalti et al. (2024c,b) proposes a novel framework that interpolates between rested and restless MABs, still assuming the rising concave condition.

3.3.4. Lower Bounds

In this section, we analyze the statistical complexity of the learning problem in both the rising and rising concave settings. To this end, we provide a regret lower bound suffered by any deterministic policy π on a class of instances which are rising and rising concave, respectively.[19] In particular, we show that rising MABs are not easier than restless MABs with bounded variation (Besbes et al., 2014, Thm. 1) and that rising concave MABs are harder than stationary MABs (Lattimore and Szepesvári, 2020, Thm. 15.2). The analysis is carried out as follows. We develop a *general recipe* for regret lower bound construction on a richer class of restless MABs, described in Section 3.3.4. Then we specialize it to both the settings of interest (Sections 3.3.4 and 3.3.4).

General Recipe for the Lower Bound

We consider a class of restless MABs with the following structure. The set of rounds $\mathbb{N}_{\geqslant 1}$ is split in windows. Let $(D_w)_{w \in \mathbb{N}_{\geqslant 1}}$ where $D_w \in \mathbb{N}_{\geqslant 1}$ be a sequence of *window widths*. A window consists of a set of rounds $\{s_w, \dots, e_w\} \subset \mathbb{N}_{\geqslant 1}$ where $s_w \coloneqq \sum_{l=1}^{w-1} D_l + 1$ and $e_w \coloneqq \sum_{l=1}^{w} D_l$, for $w \in \mathbb{N}_{\geqslant 1}$. For each window index $w \in \mathbb{N}_{\geqslant 1}$, we define two functions $\overline{\mu}_w, \widetilde{\mu}_w : [D_w] \to [0, 1]$, which we call *base* and *modified trend* respectively, that describe how the expected rewards of the arms evolve in $\{s_w, \dots, e_w\}$. In particular, in each window, *at most* one arm among the k has expected reward that follows the modified trend, while all the others have expected rewards that follow the base trend. The arm whose expected reward follows the modified trend can change between windows. We further enforce $\overline{\mu}_w(t) \leqslant \widetilde{\mu}_w(t)$ for all $w \in \mathbb{N}_{\geqslant 1}$, $t \in [D_w]$,[20] so that the arm whose expected reward follows the modified trend is the optimal one. More formally, let $w(t) \coloneqq \min\{w \in \mathbb{N}_{\geqslant 1} \text{ s.t. } e_w \geqslant t\}$ be the index of the window which contains the round $t \in \mathbb{N}_{\geqslant 1}$. For each sequence $\boldsymbol{o} = (o_w)_{w \in \mathbb{N}_{\geqslant 1}}$ with $o_w \in \{0, \dots, k\}$ in each window of index w, we define

[19] Since we are considering stochastic bandits, our lower bounds can be generalized to stochastic policies, yielding analogous results, at the cost of additional notational complexity.

[20] We consider $\overline{\mu}_w(t)$ and $\widetilde{\mu}_w(t)$ both in the domain $t \in [D_w]$ instead of in the domain $\{s_w, \dots, e_w\}$, for the sake of simplicity in the notation, as every window is defined independently from the others.

an instance $\boldsymbol{\nu}_{\boldsymbol{o}}^{\sigma} = (\nu_{\boldsymbol{o},i}^{\sigma})_{i\in[k]}$ as follows:

$$\nu_{\boldsymbol{o},i}^{\sigma}(t) := \begin{cases} 2\sigma \cdot \mathrm{Be}\left(\frac{\overline{\mu}_{w(t)}(t-s_{w(t)}+1)}{2\sigma}\right) & \text{if } i \neq o_{w(t)} \\ 2\sigma \cdot \mathrm{Be}\left(\frac{\widetilde{\mu}_{w(t)}(t-s_{w(t)}+1)}{2\sigma}\right) & \text{if } i = o_{w(t)} \end{cases}, \tag{3.21}$$

where $\sigma \geqslant 1/2$ is a constant and $\mathrm{Be}(x)$ denotes the Bernoulli distribution with parameter $x \in [0,1]$.[21] First of all observe that $\sigma \geqslant 1/2$, $\mu \in [0,1]$ imply $\mu/(2\sigma) \in [0,1]$, so that the distributions in Equation (3.21) are well-defined. Furthermore, if $X \sim 2\sigma \cdot \mathrm{Be}(\mu/(2\sigma))$, then, in virtue of Hoeffding's lemma, X is σ-subgaussian, and, by direct caclulation, it has expected value equal to μ. Thus, if $o_w = 0$, all the arms follow the base trend, otherwise, o_w corresponds to the only arm following the modified trend. We denote with $\overline{\boldsymbol{\mu}} = (\overline{\mu}_w)_{w\in\mathbb{N}_{\geqslant 1}}$ and $\widetilde{\boldsymbol{\mu}} = (\widetilde{\mu}_w)_{w\in\mathbb{N}_{\geqslant 1}}$ the sequences of base and modified trends respectively, and with $\mathcal{E}_{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}}}^{\sigma} = \{\boldsymbol{\nu}_{\boldsymbol{o}}^{\sigma} \text{ s.t. } \boldsymbol{o} \in \{0,\dots,k\}^{\mathbb{N}_{\geqslant 1}}\}$ the class of instances that they induce by varying the sequence $\boldsymbol{o}$ of optimal arms in each window. The following result, whose proof is deferred to Appendix C.1, holds.

Lemma 17 (General Lower Bound)**.** *Under the assumption that $\overline{\mu}_w(t) \leqslant \widetilde{\mu}_w(t)$ for all $w \in \mathbb{N}_{\geqslant 1}$, $t \in [D_w]$, for any deterministic policy π and learning horizon $T \in \mathbb{N}_{\geqslant 1}$, assuming Bernoulli-distributed rewards, it holds that:*

$$\sup_{\boldsymbol{\nu}\in\mathcal{E}_{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}}}} R_{\boldsymbol{\nu}}(\pi,T) \geqslant \sum_{w=1}^{w(T)} \left(1 - \frac{1}{k} - \frac{1}{\sqrt{2k}}\sqrt{\ln(2) D_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}}\right) A_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}, \tag{3.22}$$

where:

$$D_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T} := \sum_{t=s_w}^{\min\{e_w,T\}} \mathrm{D}_{\mathrm{KL}}(\overline{\mu}_w(t-s_w+1) \| \widetilde{\mu}_w(t-s_w+1)),$$

$$A_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T} := \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)),$$

for all $w \in [w(T)]$, where $\mathrm{D}_{\mathrm{KL}}(x_1\|x_2)$ for $x_1, x_2 \in [0,1]$ is the Kullback-Leibler divergence of the p.d.f. of two Bernoulli (formally defined in Appendix C.1).

This result highlights the trade-off in designing a "challenging" restless instance. On the one hand, we do not want to make the base and modified trends too far apart, otherwise it would be easy for the agent to discern one from the other. This is reflected in Equation (3.22), as the term $D_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}$ increases when the two trends diverge and contributes to reducing the regret lower bound since $A_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}$ is non-negative by construction. On the other hand, we want to maximize the area $A_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}$ between the two trends. In this way, under the assumption that $D_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}$ is small enough so

[21] For $c \in \mathbb{R}$ and $\nu \in D(\mathbb{R})$, the notation $c \cdot \nu$ denotes the probability distribution such that if $X \sim c \cdot \nu$, then $X = cY$ with $Y \sim \nu$.

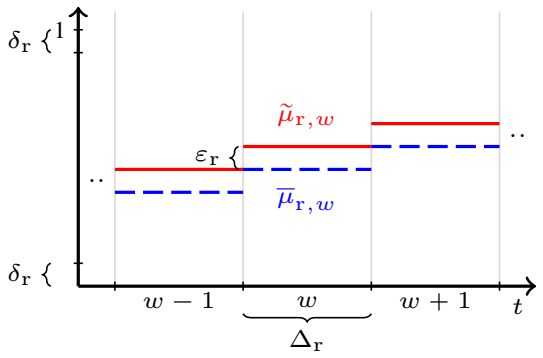


Figure 3.5: Base (dashed) and modified (solid) trends of the lower bound instances for the *rising* setting.

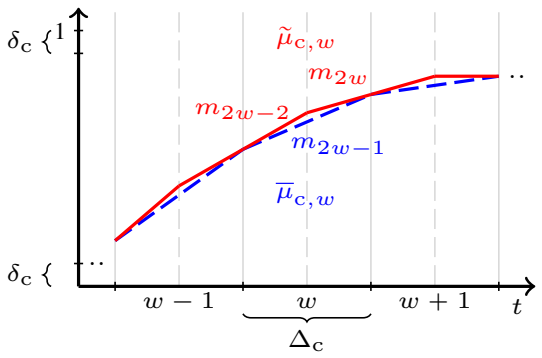


Figure 3.6: Base (dashed) and modified (solid) trends of the lower bound instances for the *rising concave* setting.

that the factor that multiplies $A_w^{\overline{\mu},\widetilde{\mu},T}$ is non-negative, we increase the regret lower bound.

Specializing the Lower Bound for the Rising Setting

In this part, we apply Lemma 17 to provide a regret lower bound for the class $\mathcal{E}_{\mathrm{r}}(T, V_T)$, holding for any deterministic policy π. To this end, we construct sequences of window widths $(D_{\mathrm{r},w})_{w\in\mathbb{N}_{\geqslant 1}}$ and of base and modified trends $\overline{\boldsymbol{\mu}}_{\mathrm{r}}$, $\widetilde{\boldsymbol{\mu}}_{\mathrm{r}}$ such that $\mathcal{E}_{\overline{\boldsymbol{\mu}}_{\mathrm{r}},\widetilde{\boldsymbol{\mu}}_{\mathrm{r}}} \subseteq \mathcal{E}_{\mathrm{r}}(T, V_T)$. A representation of the structure of the instances is depicted in Figure 3.5. We choose windows of the same width. In each window, the base and modified trend are both constant, the latter is greater than the former by a quantity $\varepsilon_{\mathrm{r}} > 0$ and the value of the modified trend in a window corresponds to the value of the base trend in the next window. In this way, we guarantee that the instances are rising no matter which arm follows the modified trend. In Appendix C.1, we formalize the instances and we prove that the following holds.

Theorem 18 (Lower Bound for the Rising Setting)**.** *For any deterministic policy π and learning horizon $T \in \mathbb{N}_{\geqslant 1}$, $T \geqslant 2^{-3}k\min\{1, V_T\}^{-2}$, assuming Bernoulli-distributed rewards, it holds that:*

$$\sup_{\boldsymbol{\nu}\in\mathcal{E}_{\mathrm{r}}(T,V_T)} R_{\boldsymbol{\nu}}(\pi, T) \geqslant \frac{1}{80} T^{\frac{2}{3}} k^{\frac{1}{3}} \min\{1, V_T\}^{\frac{1}{3}}.$$

The orders of growth for T, k, and V_T in this result match the upper bound for the general restless case with bounded variation (Besbes et al., 2014, Thm. 2) when $V_T \leqslant 1$.[22] This implies that rising MABs are not easier than general restless MABs with bounded variation despite the additional assumption. Thus, the characterization of the statistical complexity of this setting is completed.

[22] We believe this is an artifact of the analysis since, in our the lower bound construction, we have $\Upsilon(1, T) \leqslant 1$.

Specializing the Lower Bound for the Rising Concave Setting

In this part, we provide a regret lower bound for the class $\mathcal{E}_{\mathrm{c}}(T, V_T)$ holding for any deterministic policy π. In analogy to Section 3.3.4 for the rising setting, we construct sequences of window widths $(D_{\mathrm{c},w})_{w\in\mathbb{N}_{\geqslant 1}}$ and of base and modified trends $\overline{\boldsymbol{\mu}}_{\mathrm{c}}$, $\widetilde{\boldsymbol{\mu}}_{\mathrm{c}}$ such that $\mathcal{E}_{\overline{\boldsymbol{\mu}}_{\mathrm{c}},\widetilde{\boldsymbol{\mu}}_{\mathrm{c}}} \subseteq \mathcal{E}_{\mathrm{c}}(T, V_T)$. A representation of the instances is depicted in Figure 3.6. We choose again windows of the same width. In each window, the base and modified trends share the same starting and ending values. Furthermore, the end value of expected rewards in a window matches the start value of expected rewards in the next window. The end value is greater than the start value to guarantee that the instances are rising. The base trend joins the two endpoints of the expected rewards of each window with a single segment, while the modified trend uses two segments. At the beginning, it rises with a slope greater than that of the base trend until half the window. At this point, the distance between the base and the modified trend in the window is maximum. Then, the modified trend keeps rising, but with a slope that is smaller than that of the base trend, until the two trends meet at the end of the window. The pattern repeats and the slopes are chosen in such a way that the slope of the second part of the modified trend in a window (which is the smallest slope in a window) corresponds to the slope of the first part of the modified trend in the next window (which is the greatest slope of an expected reward in a window). In this way, we guarantee that the instances are rising and concave, no matter the choice of which arm follows the modified trend. In Appendix C.1, we formally present the instances and we prove the following result.

Theorem 19 (Lower Bound for the Rising Concave Setting)**.** *For any deterministic policy π and learning horizon $T \in \mathbb{N}_{\geqslant 1}$, $T \geqslant 2^{10} k \min\{1, V_T\}^{-2}$, with Bernoulli-distributed rewards, it holds:*

$$\sup_{\boldsymbol{\nu}\in\mathcal{E}_{\mathrm{c}}(T,V_T)} R_{\boldsymbol{\nu}}(\pi, T) \geqslant 2^{-15} T^{\frac{3}{5}} k^{\frac{2}{5}} \min\{1, V_T\}^{\frac{1}{5}}.$$

This result proves that regret minimization in rising concave MABs represents a harder learning problem w.r.t. stationary MABs which are characterized by the usual $\Omega(T^{1/2})$ lower bound.

3.3.5. Upper Bound for the Rising Concave Setting

In this section, we present a novel regret minimization algorithm, `Rising Concave Budgeted Exploration` (`RC-BE`(α)), designed for rising concave MABs (Algorithm 9), and analyze its performance by providing an upper bound of the expected cumulative regret suffered on a generic instance $\boldsymbol{\nu} \in \mathcal{E}_{\mathrm{c}}(T, V_T)$. We show that this upper bound attains a strictly smaller rate w.r.t. the lower bound on the expected cumulative regret on a generic restless MAB with bounded variation (Besbes et al., 2014), and thus that rising concave MABs are indeed an easier setting w.r.t. them.

Algorithm. RC-BE(α) is an improvement of the Budgeted Exploration (BE) algorithm (Jia et al., 2023), originally designed for 2-armed general restless bandits.[23] The original BE algorithm works as follows. The learning horizon T is split in windows of $D \in \mathbb{N}_{\geqslant 1}$ rounds each. In each window, the algorithm restarts. At the beginning of each window, the agent carries out an exploration phase which consists of several *round-robin* cycles. In particular, the agent keeps track of the arms alive in the current window in a set $\mathcal{A} \subseteq [k]$, initialized to $[k]$ at the beginning of each window, and, in each round-robin cycle, pulls each of these arms once. The agent cumulates the observed rewards for each arm in the variables $\hat{S}_i$ with $i \in [k]$. At the end of each round-robin cycle, the agent compares the cumulative reward of each alive arm with the maximum cumulative reward among alive arms $\hat{S}^* := \max_{i \in \mathcal{A}} \hat{S}_i$. If for $i \in \mathcal{A}$ we have $\hat{S}_i + B < \hat{S}^*$, where $B > 0$ is a parameter of the algorithm, we say that arm i has run out of budget and the agent removes it from the set of alive arms. It can happen that, after several round-robin cycles, the set of alive arms becomes a singleton: $\mathcal{A} = \{\hat{i}^*\}$. In this case, no more eliminations can happen and the agent will commit to the remaining arm $\hat{i}^*$.

RC-BE(α) extends the original algorithm as follows. It exploits the concavity of the instance through increasing window widths $D_w^{(\alpha)} := \lceil w^\alpha \rceil$ and corresponding budgets $B_w^{(\alpha)} := 2(1 + 2\sigma(D_w^{(\alpha)} \ln(2kD_w^{(\alpha)}))^{1/2})$. The rationale is the following. The algorithm suffers a high regret in windows during which the optimal arm changes. Indeed, in windows where no change happens, the algorithm is likely to commit to the best arm, suffering no regret after the initial exploration phase. Conversely, in windows where the optimal arm changes, the algorithm could commit to an arm that then becomes suboptimal, or it could fail in estimating the optimal arm. In this case, the regret increases with the distance of the expected rewards of $\hat{i}^*$ and the actual optimal arm in round t: $i_t^* \in \arg\max_{i \in [k]} \mu_i(t)$. Thanks to the concavity, the maximum increment $\max_{i \in [k]} \gamma_i(t)$ decreases as t increases. Thus, as time passes, if the optimal arm changes, it takes longer for the expected rewards of $\hat{i}^*$ and i_t^* to diverge significantly. Hence, we can restart the algorithm with a lower frequency, which is equivalent to having windows with increasing width.

Regret Analysis. RC-BE(α) partitions the set of rounds $\mathbb{N}_{\geqslant 1}$ in windows $\{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$ with $s_w^{(\alpha)} := \sum_{l=1}^{w-1} D_l^{(\alpha)} + 1$ and $e_w^{(\alpha)} := \sum_{l=1}^{w} D_l^{(\alpha)}$, for $w \in \mathbb{N}_{\geqslant 1}$. Let $w^{(\alpha)}(t) = \min\{w \in \mathbb{N}_{\geqslant 1} \text{ s.t. } e_w^{(\alpha)} \geqslant t\}$ be the index of the window that contains the round $t \in \mathbb{N}_{\geqslant 1}$. Thus, the learning horizon T is split in $w^{(\alpha)}(T)$ windows. In what follows, we bound the regret suffered by RC-BE(α) on set of windows $\mathcal{W}$ which enjoy certain properties that we introduce later. To this

[23] The extension of BE to k-armed bandits is proposed in the unpublished preprint (Jia et al., 2024) for the case of *smooth* MABs. However, we have found soundness issues in the analysis proposed there (see Appendix C.5). For this reason, we will develop an independent analysis which overcomes these issues.

Algorithm 9: RC-BE(α)

Input: $\alpha \geqslant 1$, $k \in \mathbb{N}_{\geqslant 2}$

Initialize $w \leftarrow 1$, $d \leftarrow 1$, $\mathcal{A} \leftarrow [k]$, $\mathcal{B} \leftarrow \mathcal{A}$, $\hat{S}_i \leftarrow 0$ for all $i \in [k]$
for $t \in [T]$ **do**
  if $d = D_w^{(\alpha)} + 1$ **then**
    $w \leftarrow w + 1$
    $d \leftarrow 1$
    $\mathcal{A} \leftarrow [k]$
    $\mathcal{B} \leftarrow \mathcal{A}$
    $\hat{S}_i \leftarrow 0$, for all $i \in [k]$
  end
  Pull $I_t \in \mathcal{B}$
  $\mathcal{B} \leftarrow \mathcal{B} \backslash \{I_t\}$
  Observe $R_t = X_{I_t,t}$
  $\hat{S}_{I_t} \leftarrow \hat{S}_{I_t} + R_t$
  if $\mathcal{B} = \varnothing$ **then**
    $\hat{S}^* \leftarrow \max_{i \in \mathcal{A}} \hat{S}_i$
    for $i \in [k]$ **do**
      if $i \in \mathcal{A}$ ***and*** $\hat{S}_i + B_w^{(\alpha)} < \hat{S}^*$ **then**
        $\mathcal{A} \leftarrow \mathcal{A} \backslash \{i\}$
      end
    end
    $\mathcal{B} \leftarrow \mathcal{A}$
  end
  $d \leftarrow d + 1$
end

end, we denote the regret suffered by a policy π *on a set of windows* $\mathcal{W} \subset \mathbb{N}_{\geqslant 1}$, $|\mathcal{W}| < \infty$ as:

$$R_{\boldsymbol{\nu}}(\pi, \mathcal{W}) := \sum_{w \in \mathcal{W}} \sum_{t=s_w^{(\alpha)}}^{e_w^{(\alpha)}} \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\nu}} \left[\mu_{i_t^*}(t) - \mu_{I_t}(t) \right].$$

Now, we present the properties which induce the classes of windows of interest for the analysis. In particular, we need to formally characterize the fact that, in a window, the optimal arm can change. To this end, we introduce the following definitions, in analogy to what is done in (Jia et al., 2024).

Definition 3 (Overtaking). *An arm $i \in [k]$ overtakes an arm $j \in [k]$ at time $t \in \mathbb{N}_{\geqslant 2}$ if $\mu_i(t-1) \leqslant \mu_j(t-1)$ and $\mu_i(t) \geqslant \mu_j(t)$. Formally, we write $i \uparrow_t j$ (note that $i \uparrow_t i$).*

Definition 4 (Crossing). *Two arms $i, j \in [k]$ cross at time $t \in \mathbb{N}_{\geqslant 2}$, if $i \uparrow_t j$ or $j \uparrow_t i$. Formally, we write $i \times_t j$ (note that $i \times_t i$).*

We introduce a binary relation for arms that cross in the w-th window. For $w \in \mathbb{N}_{\geqslant 1}$, $i, j \in [k]$:

$$i \,{}_w\!\times j \quad \text{iff} \quad i \times_t j \text{ for some } t \in \{s_w^{(\alpha)} + 1, \dots, e_w^{(\alpha)}\}.$$

Let ${}_w\!\times^+$ be the transitive closure of ${}_w\!\times$. ${}_w\!\times^+$ is an equivalence relation since ${}_w\!\times$ is reflexive and symmetric. For an arm $i \in [k]$, we denote with $[i]_{{}_w\times^+}$ its equivalence class w.r.t. ${}_w\!\times^+$. Let:

$$\mathcal{I}_w^* := \left\{ i \in [k] \text{ s.t. there exists } t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\} \text{ with } i \in \arg\max_{j \in [k]} \mu_j(t) \right\},$$

be the set of *optimal arms in window* w. Furthermore, we define $\mathcal{I}_w^{\times} := [i_w^*]_{{}_w\times^+}$ for some $i_w^* \in \mathcal{I}_w^*$. Observe that the definition is well posed since, in virtue of Lemma 47, it does not depend on the choice of i_w^*. For $w \in \mathbb{N}_{\geqslant 1}$, $i \in [k]$, we define the *diameter* of its equivalence class w.r.t. ${}_w\!\times^+$ as

$$d_w(i) := \max_{j,k \in [i]_{{}_w\times^+},\, t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}} \left| \mu_j(t) - \mu_k(t) \right|.$$

We use the shorthand d_w^* for $d_w(i_w^*)$ where $i_w^* \in \mathcal{I}_w^*$. The following lemma decomposes the regret suffered by $\texttt{RC-BE}(\alpha)$ during the w-th window as the sum of the regret due to the exploration phase plus the regret due to the commitment phase.

Lemma 20. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $w \in \mathbb{N}_{\geqslant 1}$ we have that:*

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \{w\}) \leqslant \underbrace{3kB_w^{(\alpha)}}_{\text{Exploration}} + \underbrace{D_w^{(\alpha)} d_w^*}_{\text{Commitment}}.$$

Thus, the regret due to exploration is proportional to the budget $B_w^{(\alpha)}$, while the regret suffered during the commitment phase depends on the width of the window $D_w^{(\alpha)}$ and on the diameter d_w^* of $\mathcal{I}_w^{\times}$. In windows where the optimal arm does not change, $\mathcal{I}_w^{\times}$ is a singleton and, thus, its diameter is 0. This reflects the fact that, in such windows, the algorithm suffers only the regret due to the exploration.

We now provide an upper bound for $d_w(i)$ with $w \in \mathbb{N}_{\geqslant 1}, i \in [k]$ which exploits concavity.

Lemma 21. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $w \in \mathbb{N}_{\geqslant 1}$, $i \in [k]$, we have that:*

$$d_w(i) \leqslant 8(1+\alpha)\left(\left|[i]_{{}_w\times^+}\right| - 1\right) w^{-1} \Upsilon_{\boldsymbol{\nu}}(1, e_w^{(\alpha)}) \leqslant 16 \alpha k w^{-1} \Upsilon_{\boldsymbol{\nu}}(1, e_w^{(\alpha)}).$$

Recall that $\Upsilon_{\boldsymbol{\nu}}(1, e_w^{(\alpha)})$ is upper bounded by k. Thus, as expected, eventually the upper bound of the diameter decreases as w increases. This reflects what we informally stated before. As time goes, due to the concavity, it takes more time for the expected rewards of arms which have

crossed to diverge significantly. Thus, it makes sense to increase the width of the windows over time.

We now discriminate between two kinds of windows: those in which the expected rewards of arms which cross (and thus of the arms which belong to $\mathcal{I}_w^{\times}$) do not diverge significantly and those in which, instead, the converse happens. More formally, let $d \in (0, k]$:

$$\mathcal{W}_{\leqslant d}(T) \coloneqq \left\{w \in [w^{(\alpha)}(T)] \text{ s.t. } d_w(i) \leqslant d \text{ for all } i \in [k]\right\},$$

$$\mathcal{W}_{>d}(T) \coloneqq \left\{w \in [w^{(\alpha)}(T)] \text{ s.t. } d_w(i) > d \text{ for some } i \in [k]\right\}.$$

In the second class of windows, we have no upper bound to the diameter d_w^* other than that of Lemma 21, which considers a worst-case scenario in which the divergence of the expected rewards of the arms which cross is the maximum possible. We now show that this scenario, in the rising concave setting, can happen only a limited number of times. In particular, this is translated into an upper bound to the number of windows in $\mathcal{W}_{>d}(T)$, which is captured by the following lemma.

Lemma 22. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $T \in \mathbb{N}_{\geqslant 1}$, $d \in (0, k]$, we have that:*

$$|\mathcal{W}_{>d}(T)| \leqslant 9 \ln \left(3e^{(\alpha)}_{w^{(\alpha)}(T)} k/d\right) k^{\frac{5}{2}} d^{-\frac{1}{2}}.$$

Informally, this lemma states that, in the rising concave setting, it cannot happen in too many windows that the expected rewards of arms which cross diverge significantly (*i.e.*, more than d).

We use this fact to conclude the analysis. In particular, observe that we can always upper bound the regret suffered on a set of windows $\mathcal{W}$ as $R_{\boldsymbol{\nu}}(\pi, \mathcal{W}) \leqslant |\mathcal{W}| \max_{w \in \mathcal{W}} R_{\boldsymbol{\nu}}(\pi, \{w\})$. We use this to upper bound the regret suffered on both $\mathcal{W}_{\leqslant d}(T)$ and $\mathcal{W}_{>d}(T)$. In the first case, we observe that $|\mathcal{W}_{\leqslant d}(T)| \leqslant w^{(\alpha)}(T)$ and use the definition of $\mathcal{W}_{\leqslant d}(T)$ together with Lemma 20 to bound $\max_{w \in \mathcal{W}_{\leqslant d}(T)} R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \{w\})$. In the second case, we use Lemma 22 to upper bound $|\mathcal{W}_{>d}(T)|$ and Lemma 20 together with Lemma 21 to deal with $\max_{w \in \mathcal{W}_{>d}(T)} R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \{w\})$. These observations lead to the following result which is formally proven in Appendix C.2.

Theorem 23 (Upper Bound for the Rising Concave Setting)**.** *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $T \in \mathbb{N}_{\geqslant 24}$, we have that:*

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), T) \leqslant 2^{15} \alpha^3 (1+\sigma) \left(\ln\left(\alpha k T^3\right)\right)^{\frac{3}{2}} \left(k^3 T^{\frac{3/4\alpha}{1+\alpha}} + k^3 T^{\frac{5/4\alpha - 1}{1+\alpha}} \Upsilon_{\boldsymbol{\nu}}(1, T) + k T^{\frac{1+\alpha/2}{1+\alpha}}\right).$$

In particular, for $\alpha' := 8/3$*, we get:*

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha'), T) = \widetilde{\mathcal{O}}\left(k^3 T^{\frac{6}{11}} + k^3 T^{\frac{7}{11}} \Upsilon_{\boldsymbol{\nu}}(1, T) + k T^{\frac{7}{11}}\right).$$

Furthermore, for $\alpha'' := (8 - 8\log_T(k\sqrt{V_T}))/(3 + 8\log_T(k\sqrt{V_T}))$*, under the additional assumptions* $\boldsymbol{\nu} \in \mathcal{E}_{\mathrm{c}}(T, V_T)$*,* $T \geqslant \max\{k^{-8/3} V_T^{-4/3} + 1, k^{16/5} V_T^{8/5}\}$*, we get:*

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha''), T) = \widetilde{\mathcal{O}}\left(k^{\frac{27}{11}} T^{\frac{6}{11}} V_T^{-\frac{3}{11}} + k^{\frac{15}{11}} T^{\frac{7}{11}} V_T^{\frac{2}{11}}\right).$$

By looking at the algorithm and at Theorem 23, we observe how by selecting $\alpha = 8/3$, we achieve a regret of order $\widetilde{\mathcal{O}}(T^{7/11})$ without the knowledge of either the total variation V_T or the learning horizon T, making it an anytime algorithm, at the price of worst dependence on k and V_T. This result shows that the regret minimization problem in rising concave MABs is indeed easier w.r.t. general restless MABs with bounded variation (Besbes et al., 2014) and rising MABs. Indeed, the regret $\widetilde{\mathcal{O}}(T^{7/11})$ in our upper bound is smaller than that of the lower bound for restless MABs with bounded variation (Besbes et al., 2014, Theorem 1) and rising MABs (Theorem 18), i.e., $\Omega(T^{2/3})$.

3.3.6. Numerical Simulations

In this section, we present the results of numerical simulation of $\texttt{RC-BE}(\alpha)$ compared to state-of-the-art algorithms for restless, restless rising concave, and stationary MABs.[24]

Baselines. We consider the baseline algorithms: `Rexp3` (Besbes et al., 2014), an algorithm for restless MABs based on a variation budget; `R-less-UCB` (Metelli et al., 2022), an algorithm for restless rising concave MABs; and `UCB1` (Auer et al., 2002a), one of the most effective algorithms for stationary MABs. The choices of the parameters of the algorithms are reported in Appendix C.4.

Setting. The algorithms are evaluated for $T = 5 \cdot 10^6$ rounds on synthetic instances with $k = 5$ arms. The stochasticity is realized by adding Gaussian noise with standard deviation $\sigma = 0.1$. The curves of the expected rewards have the functional form $f(t) = c(1 - \exp(-sat/T))$ for $t \in [T]$ where $a, c \in (0, 1]$, $s = 50$, and are reported in Figure 3.7. We compare the algorithms in terms of empirical cumulative regret $\widehat{R}_{\boldsymbol{\nu}}(\pi, t)$ which is the empirical counterpart of the expected cumulative regret $R_{\boldsymbol{\nu}}(\pi, t)$ at round t averaged over multiple independent runs. In each simulation, the parameter α of $\texttt{RC-BE}(\alpha)$ is set to $\alpha = 8/3$, as suggested by Theorem 23.

[24] Additional simulations are reported in Appendix C.4. The code to reproduce the results is available in the supplementary material.

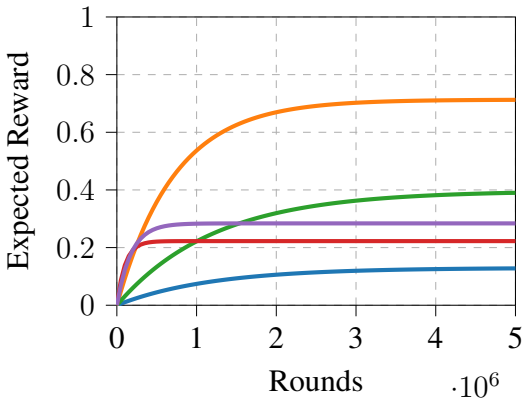


Figure 3.7: (a) Expected rewards.

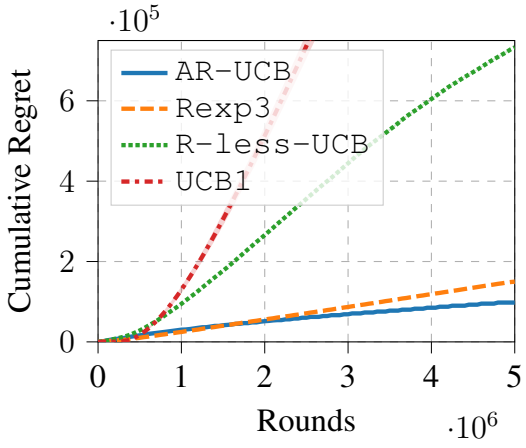


Figure 3.8: (b) Cumulative regret (10 runs, mean $\pm$ std).

Figure 3.9: Instance and results of the experimental validation.

Results. The empirical cumulative regret suffered by the algorithms is shown in Figure 3.8. We observe that `RC-BE`(α) is the algorithm that achieves the lowest regret at the horizon. `UCB1` has the lowest regret in the first rounds, afterwards its regret starts increasing when the optimal arm changes. This is consistent with the fact that we are violating the stationarity assumption on which the algorithm relies. `Rexp3` is an algorithm which restarts at a fixed frequency. In particular, the number of restarts has order $T^{1/3}$. Thus, in this simulation, there are $\approx 10^2$ restarts, and, by looking at the figure, it is not possible to appreciate the behavior of the algorithm between one restart and the next. For this reason, `Rexp3` shows a cumulative regret which increases linearly. This is consistent with the fact that the algorithm is not anytime. `R-less-UCB`, consistently with its theoretical guarantees, shows a sublinear growth of the cumulative regret. Its estimator relies on a rested model of the evolution of the expected rewards of the arms, penalizing the empirical performance.

3.3.7. Discussion and Future Directions

In this paper, we studied the restless rising and rising concave MABs, where the expected rewards of the arms are non-decreasing and non-decreasing concave in the number of played rounds, respectively. We derived lower bounds to the expected cumulative regret in both settings. The lower bound in the rising setting has order $\Omega(T^{2/3})$ and implies that the non-decreasing expected reward assumption does not simplify the learning problem w.r.t. the general restless setting with bounded variation, and so that all the algorithms which are optimal for the general setting are optimal also in this special subclass, closing in this way the gap present in the literature.

Thus, for the rising setting, we provided a positive answer to our **Question 1** and a negative answer to our **Question 2**. The lower bound in the rising concave setting has order $\Omega(T^{3/5})$ and implies that rising concave MABs represent a statistically harder problem w.r.t. stationary MABs. After having presented two statistical barriers for these settings, we developed a learning algorithm with the goal of exploiting the more structured model of rising concave MABs. To this end, we designed $\texttt{RC-BE}(\alpha)$, and we derived an upper bound to its expected regret of order $\widetilde{\mathcal{O}}(T^{7/11})$. This result implies that the non-decreasing expected reward assumption, together with the concave expected reward assumption, simplifies the learning problem w.r.t. the same setting without concavity. Thus, for the rising concave setting, we provided a positive answer for both **Question 1** and **Question 2**. The natural future research direction includes closing the gap in rising concave MABs which is now only $7/11 - 3/5 = 2/55$ in the exponent of T.

4 | Heavy-Tailed Bandits

Heavy-tailed distributions naturally arise in several settings, from finance to telecommunications. While regret minimization in MABs under subgaussian or bounded rewards has been widely studied, learning with heavy-tailed distributions only gained popularity over the last decade. In this chapter, we consider the setting in which the reward distributions have finite absolute raw moments of maximum order $1 + \epsilon$, uniformly bounded by a constant $u < +\infty$, for some $\epsilon \in (0, 1]$. This setting represents a generalization of the one introduced in Chapter 2, as heavy-tailed distribution are way less well-behaved rather than sub-Gaussian distributions. Our contribution is twofold: in the first section, we discuss the problem of regret minimization in heavy-tailed MABs when u and ϵ are both unknown, a though open-problem to which we partially found a solution; in the second section, we jointly tackle the problem of heavy-tailedness of the rewards and non-stationarity of the environment.

4.1. Adaptation to Unknown Distributional Parameters in Heavy-Tailed Bandits

In this section, we study the regret minimization problem in heavy-tailed MABs when ϵ and u are unknown to the learner and it has to adapt. First, we show that adaptation comes at a cost and derive two negative results proving that the same regret guarantees of the non-adaptive case cannot be achieved with no further assumptions. Then, we devise and analyze a fully data-driven trimmed mean estimator and propose a novel adaptive regret minimization algorithm, `AdaR-UCB`, that leverages such an estimator. Finally, we show that `AdaR-UCB` is the first algorithm that, under a known distributional assumption, enjoys regret guarantees nearly matching those of the non-adaptive heavy-tailed case. This section presents Genalti et al. (2024a), a joint project with Lupo Marsigli, Nicola Gatti and Alberto Maria Metelli, published at the *Annual Conference on Learning Theory (COLT)*.

4.1.1. Introduction

We investigate the stochastic MAB problem under the assumption of *heavy-tailed* (HT) reward distributions, introduced in the seminal work Bubeck et al. (2013a). We follow the same notation introduced in Chapter 2 Although most of the literature in stochastic MABs usually assumes a convenient tail property of the reward distributions, like *subgaussian* (Lattimore and Szepesvári, 2020) or *bounded* (Auer et al., 2002a) rewards, in many practical scenarios, such as financial environments (Gagliolo and Schmidhuber, 2011) or network routing problems (Liebeherr et al., 2012), where uncertainty has a significant impact, heavy-tailed distributions naturally arise. In these cases, the tails decay more slowly than a Gaussian, the moment-generating function is no longer finite, and the moments of all finite orders might not exist. This prevents the application of standard concentration tools, such as Hoeffding's inequality (Boucheron et al., 2013), calling for more complex technical tools.

We assume that the *absolute raw moments* of the reward distributions of order up to $1+\epsilon$, with $\epsilon \in (0,1]$ (*i.e.,* moment order) are finite and uniformly bounded by a constant $u \in \mathbb{R}_{\geqslant 0}$ (*i.e.,* moment bound), namely:

$$\boldsymbol{\nu} \in \mathcal{P}_{\text{HT}}(\epsilon, u)^k := \left\{ \boldsymbol{\nu} \in \Delta(\mathbb{R})^k \, : \, \mathop{\mathbb{E}}_{X \sim \nu_i} \left[|X|^{1+\epsilon} \right] \leqslant u, \;\; \forall i \in [k] \right\}, \tag{4.1}$$

In Bubeck et al. (2013a), the authors assume that ϵ *and* u *are known to the learner*. They show that if the variance is finite (*i.e.*, $\epsilon = 1$), but the higher order moments are not, the same (apart from constants) instance-dependent regret guarantees of order $O\big(\sum_{i:\Delta_i>0} \frac{\sigma^2}{\Delta_i} \ln T\big)$ attained in the subgaussian setting (Lattimore and Szepesvári, 2020) can be achieved. However, for general $\epsilon \in (0,1)$, the instance-dependent regret becomes of order $O\big(\sum_{i:\Delta_i>0} \big(\frac{u}{\Delta_i}\big)^{1/\epsilon} \ln T\big)$, showing the detrimental effect of ϵ on the dependence of the suboptimality gaps. Moreover, they show that these regret guarantees are tight (up to constant terms) by deriving the corresponding asymptotic lower bound. From a worst-case regret perspective, the presented results translate into a regret bound of order $\widetilde{O}\big(k^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}}\big)$, for sufficiently large T, matching the lower bound, up to logarithmic terms. This regret bound degenerates to linear when $\epsilon \to 0$, *i.e.,* when only absolute moment of order 1 exists. However, these matching results are obtained thanks to the knowledge of both ϵ and u, *i.e.,* by *non-adaptive* algorithms. Indeed, ϵ and u are needed by the *algorithm* to drive exploration via the optimistic index, and, in some cases, to construct the expected reward *estimator* too. Nevertheless, recent works have shed light on the possibility of removing this knowledge at the cost of additional assumptions (e.g., Lee et al., 2020; Ashutosh et al., 2021; Huang et al., 2022). In particular, Huang et al. (2022) introduce the *truncated non-negativity* assumption designed for losses, that, by converting it for rewards, leads to the following *truncated non-positivity* assumption.

Assumption 6 (Truncated Non-Positivity)**.** *Let $\boldsymbol{\nu}$ be a bandit. For the optimal arm i^*, we have:*

$$\mathbb{E}_{X\sim\nu_{i^*}}\left[X\mathbb{1}_{\{|X|>M\}}\right]\leqslant 0,\quad \forall M\geqslant 0. \tag{4.2}$$

This assumption requires that the optimal arm 1 has a larger probability mass on the negative semi-axis but still allows the distribution to have an arbitrary support covering, potentially, the whole $\mathbb{R}$. To the best of the authors' knowledge, this is the only assumption in literature truly independent of the values of ϵ and u. Additionally, as discussed by Huang et al. (2022), it is relatively weak if compared to other standard assumptions in the bandit literature. Under Assumption 6, *without the knowledge of ϵ and u*, Huang et al. (2022) provide an (ϵ,u)-*adaptive*[1] regret minimization algorithm, `AdaTINF`, that succeeds in matching the worst-case regret lower bound of Bubeck et al. (2013a) derived for the non-adaptive case. However, no instance-dependent analysis is provided of `AdaTINF`[2] and the following research questions remain open:

Research Question 1 *Is Assumption 6 needed to devise (ϵ,u)-adaptive algorithms (with unknown (ϵ,u)) matching the worst-case lower bound of order $\Omega\big(k^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}}\big)$ (*i.e., *as if we knew (ϵ,u))?*

Research Question 2 *Is it possible, under Assumption 6, to devise (ϵ,u)-adaptive algorithms (with unknown (ϵ,u)) matching the instance-dependent regret lower bound of order $\Omega\big(\sum_{i:\Delta_i>0}\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}\ln T\big)$ (*i.e., *as if we knew (ϵ,u))?*

Original Contributions. In this section, we investigate the regret minimization problem in heavy-tailed bandits giving up the knowledge of ϵ and u. Specifically, we address Research Question 1 and Research Question 2. The original contributions of the paper are summarized as follows:

- In Section 4.1.3, we address Research Question 1, by characterizing the challenges of the regret-minimization problem in HT bandits without knowing ϵ and u. In particular, we provide two *negative results* (Theorems 25 and 26), showing that, without any additional assumption, there exists no (ϵ,u)-adaptive algorithm that achieves the same worst-case regret guarantees as if ϵ or u were known (Bubeck et al., 2013a). These results provide a first justification of Assumption 6. Furthermore, we show how Assumption 6 does not reduce the complexity of the regret minimization problem even in the non-adaptive case (Theorem 27). These results rely on accurately defined HT bandit instances and information theory tools for deriving the lower bounds.
- In Section 4.1.4, we enhance the *trimmed mean* estimator, commonly used in HT bandits,

[1] We use the word *adaptive* to qualify algorithms that do not know the values of ϵ and/or u.

[2] Huang et al. (2022) actually provide algorithm `OptHTINF` with an instance-dependent analysis that, however, does not match the asymptotic lower bound of Bubeck et al. (2013a).

to make it fully data-driven. Indeed, in the seminal paper (Bubeck et al., 2013a), both (i) the trimming threshold and (ii) the upper confidence bound were computed thanks to the knowledge of ϵ and u. Taking inspiration from Huber regression (Wang et al., 2021), we overcome (i) the need for ϵ and u in the estimator by developing a novel approach to recover an estimated threshold via *root-finding*. Leveraging an analysis based on the *self-bounding functions* (Maurer, 2006; Maurer and Pontil, 2009), we control the accuracy of the estimated threshold (Lemma 29). In particular, we show that our threshold underestimates (in high probability) the one proposed by Bubeck et al. (2013a). Furthermore, we overcome (ii) by resorting to *empirical Bernstein inequality* (Maurer and Pontil, 2009). This way, differently from Bubeck et al. (2013a), we use the empirical variance to eliminate the dependence on ϵ and u in the upper confidence bound (Lemma 28), preserving the desirable concentration properties of the delivered estimate (Theorem 30).

- In Section 4.1.5, we address Research Question 2, by devising and analyzing a novel (ϵ, u)-adaptive regret minimization algorithm, `Adaptive Robust UCB` (`AdaR-UCB`, Algorithm 11), that operates without the knowledge of ϵ and u. `AdaR-UCB` is an *optimistic anytime* algorithm that builds upon `Robust UCB` of Bubeck et al. (2013a), leveraging our trimmed mean estimator with estimated threshold. First, we show that, under Assumption 6, `AdaR-UCB` attains an instance-dependent regret bound of order $O\big(\sum_{i:\Delta_i>0}\big(\big(\frac{u}{\Delta_i}\big)^{1/\epsilon}+\frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\big)\ln T\big)$ (Theorem 31). This result shows that `AdaR-UCB` nearly matches the instance-dependent lower bound of Bubeck et al. (2013a) for the non-adaptive case, apart from the second logarithmic term, which, however, does not depend on the reciprocal of the suboptimality gaps, and originates from an additional forced exploration needed for computing the empirical threshold.[3] Moreover, we show that `AdaR-UCB` suffers a worst-case regret bound of order $\widetilde{O}\big(k^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}}\big)$ (Theorem 32), matching, up to logarithmic terms, the minimax lower bound of the non-adaptive case (Bubeck et al., 2013a). To the best of authors' knowledge, `AdaR-UCB` is the first (ϵ, u)-adaptive algorithm for HT bandits that nearly matches both the instance-dependent and worst-case lower bounds of the non-adaptive case, under conditions (Assumption 6) not explicitly formulated in terms of ϵ and u.

In Section 4.1.2 we provide an up-to-date literature review on *adaptivity* in heavy-tailed bandits. The proofs of the results presented in the main paper are reported in Appendix D.2.

4.1.2. Related Works

During the last ten years, the stochastic heavy-tailed bandit problem has been steadily increasing in popularity. In this section, we summarize the main contributions, with a particular focus on

[3] A similar additional term appears in the instantiation of `Robust UCB` with Catoni estimator (Bubeck et al., 2013a).

Algorithm	Regret Bounds				ϵ-adaptive		u-adaptive		Assumption
	Instance-dependent	Matching?§	Worst-case	Matching?¶	Estimator	Algorithm	Estimator	Algorithm	
`Robust UCB` - TM (Bubeck et al., 2013a)	$\sum_{i:\Delta_i>0}\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}(\log T)^{\frac{\epsilon}{1+\epsilon}}$	✓	✗	✗	✗	✗	—
`Robust UCB` - MoM ⋆ (Bubeck et al., 2013a)	$\sum_{i:\Delta_i>0}\left(\frac{v}{\Delta_i}\right)^{1/\epsilon}\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}v^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}(\log T)^{\frac{\epsilon}{1+\epsilon}}$	✓	✓	✗	✓	✗	—
`Robust UCB` - Catoni ⋆ (Bubeck et al., 2013a)	$\sum_{i:\Delta_i>0}\left(\frac{v}{\Delta_i}+\Delta_i\right)\log T$	✓	$\sqrt{vKT\log T}+\sum_{i:\Delta_i>0}\Delta_i\log T$	✓	✗	✗	✓	✗	ϵ=1 only
`Robust MOSS` (Wei and Srivastava, 2020)	$\sum_{i:\Delta_i>0}\log\left(\frac{T\Delta_i^{\frac{1+\epsilon}{\epsilon}}}{K}\right)\frac{1}{\Delta_i^{1/\epsilon}}$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$	✓	✗	✗	✗	✗	—
$\texttt{KL}_{\texttt{inf}}$`-UCB` † (Agrawal et al., 2021)	$\sum_{i:\Delta_i>0}\frac{\log T}{D^{\inf}_{\mathrm{KL}}(\nu_i,\mu_1)}$	✓	—	—	✓	✗	✓	✗	—
`APE`^{2} (Lee et al., 2020)	$\sum_{i:\Delta_i>0}\left(e^u+(T\Delta_i^{\frac{1+\epsilon}{\epsilon}}\log K)^{\frac{1+\epsilon}{\epsilon\log K}}\right)\frac{1}{\Delta_i^{1/\epsilon}}$	✗	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}\log T e^u$	✗	✗	✗	✓	✓	—
`MR-APE`^{2} (Lee and Lim, 2022)	$\sum_{i:\Delta_i>0}\left(Ke^u+\log\left(\frac{T\Delta_i^{\frac{1+\epsilon}{\epsilon}}}{K}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\frac{1}{\Delta_i^{1/\epsilon}}$	✗	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}e^u$	✗	✗	✗	✓	✓	—
`HTINF` (Huang et al., 2022)	$\sum_{i:\Delta_i>0}\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$	✓	✗	✗	✗	✗	Assumption 1
`OptHTINF` (Huang et al., 2022)	$\sum_{i:\Delta_i>0}\left(\frac{u^2}{\Delta_i^{2-\epsilon}}\right)^{1/\epsilon}\log T$	✗	$K^{\frac{\epsilon}{2}}u^{\frac{1}{1+\epsilon}}T^{\frac{2-\epsilon}{2}}$	✗	✓	✓	✓	✓	Assumption 1
`AdaTINF` (Huang et al., 2022)	—	—	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$	✓	✓	✓	✓	✓	Assumption 1
`R-UCB-TEA` ‡ (Ashutosh et al., 2021)	$\sum_{i:\Delta_i>0}\frac{f(T)}{1-\frac{2}{\Delta_i\log f(t)}}\log T$	✗	—	—	✓	✓	✓	✓	T s.t. $3u\log f(T)<2f(T)^{\epsilon}$
`R-UCB-MoM` ‡ (Ashutosh et al., 2021)	$\sum_{i:\Delta_i>0}\Delta_i\left(\frac{2f(T)}{\Delta_i}\right)^{\frac{1}{g(T)}}\log T$	✗	—	—	✓	✓	✓	✓	T s.t. $\frac{g(T)<\frac{\epsilon}{1+\epsilon}}{f(T)>(12u)^{\frac{1}{1+\epsilon}}}$
`AdaR-UCB` **(ours)**	$\sum_{i:\Delta_i>0}\left(\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}+\frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\right)\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}(\log T)^{\frac{\epsilon}{1+\epsilon}}+\sum_{i:\Delta_i>0}\frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\log T$	✓	✓	✓	✓	✓	Assumption 1

⋆ The bound depends on the centered absolute moment $v := \max_{i\in[k]}\mathbb{E}_{X\sim\nu_i}[|X-\mu_i|^{1+\epsilon}]$ of order $1+\epsilon$.
† $D^{\inf}_{\mathrm{KL}}(\eta,x) := \inf\{D_{\mathrm{KL}}(\eta,\kappa) : \kappa\in\mathcal{P}_{\mathrm{HT}}(\epsilon,u) \text{ and } \mathbb{E}_{X\sim\kappa}[X]\geqslant x\}$.
‡ f and g are to be given in input. Choosing an optimal value of those would require knowing ϵ and u.
§ Matching the instance-dependent lower bound for the non-adaptive case w.r.t. T, $1/\Delta_i$, u (or v), and ϵ, up to constants.
¶ Matching worst-case lower bound for the non-adaptive case w.r.t. T, k, u (or v), and ϵ, up to logarithmic terms.

Table 4.1: Comparison with the state-of-the-art. The regret bounds are deprived by constants.

partially adaptive approaches. Table 4.1 provides a comprehensive comparison.

Bubeck et al. (2013a) represents the most influential work in this area, formally introducing the setting, deriving both instance-dependent and worst-case lower bounds, and proposing the first non-adaptive algorithm, namely `Robust UCB`. Such an algorithm can be instanced with three robust estimators: *trimmed mean* (TM), *median of means* (MoM), and *Catoni estimator* (Catoni) achieving near-optimal regret guarantees from both instance-dependent and worst-case cases. The first minimax optimal algorithm was proposed in Wei and Srivastava (2020), namely `Robust MOSS`, removing the $(\ln T)^{\frac{\epsilon}{1+\epsilon}}$. Instead, in Agrawal et al. (2021), the authors propose $\texttt{KL}_{\texttt{inf}}$`-UCB` attaining an asymptotically-optimal instance-dependent upper bound, highlighting the dependence on the instance with the KL-divergence, similarly as Garivier and Cappé (2011)

for non-heavy-tailed bandits. These algorithms, however, require the knowledge of ϵ and u, *i.e.,* they are non-adaptive.[4] In Ashutosh et al. (2021), the authors show that *adaptivity* comes at a cost in both subgaussian and heavy-tailed bandits. In particular, logarithmic instance-dependent regret is unachievable when no further information on the environment is available. They introduce two algorithms, namely `R-UCB-TEA` and `R-UCB-MoM`, exploiting the TM and the MoM estimators, respectively. Although, in principle, they do not require the knowledge of ϵ or u for execution, logarithmic regret cannot be achieved but only approached with arbitrary precision. Moreover, the bounds hold only for a learning horizon T larger than a threshold depending on ϵ and u. No worst-case analysis is presented.

The closest work to ours is Huang et al. (2022) where the authors introduce the *adversarial heavy-tailed bandits*, in which an adversary chooses HT distributions for the losses. They first introduce the *truncated non-negativity* (analogous to our Assumption 6 for rewards) representing, to the best of the authors' knowledge, the only assumption not explicitly related to ϵ and u. Three algorithms are provided: `HTINF`, `OptHTINF`, and `AdaTINF`, all analyzed under this assumption. `HTINF` requires knowledge of both ϵ and u and it is nearly optimal. Differently, both `OptHTINF` and `AdaTINF` are (ϵ, u)-adaptive. However, the instance-dependent bound of `OptHTINF` exposes an inconvenient dependence on $\left(\frac{u^2}{\Delta_i^{2-\epsilon}}\right)^{1/\epsilon}$ and the worst-case bound scales with $T^{\frac{2-\epsilon}{2}}$, both failing to match the lower bounds of the non-adaptive setting. Finally, the worst-case bound of `AdaTINF` matches the non-adaptive lower bound. However, the authors show that no logarithmic instance-dependent regret can be obtained by `AdaTINF`. Finally, Lee et al. (2020) introduces the `APE`^{2} algorithm, adaptive in u, that, unfortunately, does not achieve logarithmic instance-dependent regret that, instead, scales with $T^{\frac{1}{1+\epsilon}} \ln T$ and displays an inconvenient exponential dependence e^u. A modified version of this algorithm, namely `MR-APE`^{2}, introduced in Lee and Lim (2022), succeeds in removing the polynomial dependence on T, now poly-logarithmic $(\ln T)^{\frac{1+\epsilon}{\epsilon}}$, but maintains the dependence on e^u.

4.1.3. Minimax Lower Bounds for Adaptive Heavy-Tailed Bandits

In this section, we address Research Question 1, by analyzing the challenges of the (ϵ, u)-adaptive regret minimization problem, *i.e.,* without the knowledge of ϵ and u. We start by revising the minimax regret lower bound derived in Bubeck et al. (2013a) for the *non-adaptive* case, *i.e.,* when ϵ and u are known (Theorem 24). Then, in Section 4.1.3, we provide two novel *negative results* showing that achieving the same worst-case regret guarantees when either u (Theorem 25) or ϵ (Theorem 26) are unknown is not possible. [5] Finally, in Section 4.1.3, we derive a new

[4]The *truncated mean* requires the knowledge of ϵ and u in the construction of the expected reward estimator too.

[5]We remark that from the instance-dependent regret perspective, a negative answer to the possibility of achieving logarithmic regret with adaptive algorithms for heavy-tailed bandits has been already provided in (Ashutosh et al.,

minimax regret lower bound under the *truncated non-positivity* (Assumption 6), illustrating how, even in the non-adaptive case, this assumption does not lead to smaller regret lower bounds.

Let us start by recalling the minimax regret lower bound for the non-adaptive case.

Theorem 24 (Minimax lower bound – non-adaptive, Bubeck et al. (2013a))**.** *Fix $\epsilon \in (0, 1]$ and $u \geqslant 0$. For every algorithm π, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $k \in \mathbb{N}_{\geqslant 2}$, it holds that:*

$$\sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^k} R_{\boldsymbol{\nu},T}(\pi) \geqslant c_0 k^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}}, \tag{4.3}$$

where $c_0 > 0$ is a constant independent of u, ϵ, k, and T.

This result shows how the dependency on T deteriorates as ϵ approaches 0 and, instead, when the variance is finite, *i.e.*, $\epsilon = 1$, the lower bound displays the same order as the one for stochastic MABs with subgaussian rewards (Lattimore and Szepesvári, 2020).

Negative Results about Adaptivity

We now move to our negative results about the possibility of matching the minimax regret lower bound of the non-adaptive setting using (ϵ, u)-adaptive algorithms. The following result shows that any u-adaptive algorithm cannot achieve the same regret as in the non-adaptive case in Theorem 24.

Theorem 25 (Minimax lower bound – u-adaptive)**.** *Fix $\epsilon \in (0, 1]$. For every algorithm π, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $k \in \mathbb{N}_{\geqslant 2}$, it holds that:*

$$\sup_{u \geqslant 0} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^k} \frac{R_{\boldsymbol{\nu},T}(\pi)}{u^{\frac{1}{1+\epsilon}}} = +\infty. \tag{4.4}$$

More precisely, for every $u' \geqslant u \geqslant 0$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon, u')$ such that:

$$\max\left\{ \frac{R_{\boldsymbol{\nu},T}(\pi)}{u^{\frac{1}{1+\epsilon}}}, \frac{R_{\boldsymbol{\nu}',T}(\pi)}{(u')^{\frac{1}{1+\epsilon}}} \right\} \geqslant c_1 \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(1+\epsilon)^2}} T^{\frac{1}{1+\epsilon}}, \tag{4.5}$$

where $c_1 > 0$ is a constant independent of u, u', and T.

Some remarks are in order. First, let us observe that for proving the negative result, we have studied the ratio $R_{\boldsymbol{\nu},T}(\pi)/u^{\frac{1}{1+\epsilon}}$. Indeed, if there exists a u-adaptive regret minimization algorithm matching the lower bound for the non-adaptive case (Theorem 24), this ratio would not depend

2021, Theorem 1). Thus, our focus is on the minimax regret perspective.

on u anymore. It is convenient to start commenting on Theorem 25 from the lower bound in Equation (4.5). Here, we show the existence of two heavy-tailed bandit instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$, characterized by the same moment order ϵ but possibly different moment bounds $u' \geqslant u$, for which any algorithm suffers (apart from constants) a regret that preserves the dependence on T but introduces a dependence on the ratio u'/u. Since we can make the ratio arbitrarily large by varying $u, u' \in \mathbb{R}_{\geqslant 0}$, we conclude the statement in Equation (4.4) showing that the minimax lower bound degenerates to infinity. This shows that, with no additional assumptions, there exists no u-adaptive algorithm matching the lower bound for the non-adaptive case (Theorem 24).

We now present the counterpart negative result concerning adaptivity to the moment order ϵ.

Theorem 26 (Minimax lower bound – ϵ-adaptive)**.** *Fix $u = 1$. For every algorithm π, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $k \in \mathbb{N}_{\geqslant 0}$, it holds that:*

$$\sup_{\epsilon \in (0,1]} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^k} \frac{R_{\boldsymbol{\nu},T}(\pi)}{T^{\frac{1}{1+\epsilon}}} \geqslant c_2 T^{\frac{1}{16}}. \tag{4.6}$$

More precisely, for every $\epsilon, \epsilon' \in (0, 1]$ with $\epsilon' \leqslant \epsilon$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon', u)$ such that:

$$\max\left\{ \frac{R_{\boldsymbol{\nu},T}(\pi)}{T^{\frac{1}{1+\epsilon}}}, \frac{R_{\boldsymbol{\nu}',T}(\pi)}{T^{\frac{1}{1+\epsilon'}}} \right\} \geqslant c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon)(1+\epsilon')^2}}, \tag{4.7}$$

where $c_2 > 0$ is a constant independent of ϵ, ϵ', and T.

Differently from Theorem 25, here we target the ratio $R_{\boldsymbol{\nu},T}(\pi)/T^{\frac{1}{1+\epsilon}}$ for deriving the negative result. Indeed, having fixed $u = 1$, if an ϵ-adaptive algorithm exists matching the lower bound of Theorem 24, then, the considered ratio would not depend on T anymore. Starting from the lower bound of Equation (4.7), we observe that there exist two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$, with ϵ and ϵ' as moment orders, for which the ratio is lower bounded by a function dependent on T. Since $\epsilon \geqslant \epsilon'$, the exponent to which T is raised is non-negative and, consequently, the lower bound is a non-decreasing function of T. By letting ϵ and ϵ' range in $[0, 1)$, we obtain the minimax bound of Equation (4.6), displaying a gap of order $T^{\frac{1}{16}}$, which is attained by taking $\epsilon = 1$ and $\epsilon' = 1/3$. This result shows that there exists no ϵ-adaptive algorithm able to suffer the same regret as in the non-adaptive case of Theorem 24.

Combining Theorem 25 with Theorem 26, we conclude the non-existence of an (ϵ, u)-adaptive algorithm suffering the same regret guarantees as in the non-adaptive case. It is worth noting that the constructions employed for deriving the lower bounds presented in this section violate Assumption 6.

Minimax Lower Bound under Assumption 6

The results presented above show that, if our goal is to match the worst-case bound of the non-adaptive case of Theorem 24, we surely need to enforce additional assumptions. Huang et al. (2022) succeeds in this task by using the *truncated non-positivity* assumption (or more precisely, its dual version for losses). We may wonder whether enforcing Assumption 6 radically simplifies the problem. In the following, we show that this is not the case, by deriving a novel minimax lower bound for the non-adaptive case under this assumption of the same order as that of Theorem 24.

Theorem 27 (Minimax lower bound under Assumption 6 - non-adaptive)**.** *Fix $\epsilon \in (0, 1]$ and $u \geqslant 0$. For every algorithm π, sufficiently large learning horizon $T \in \mathbb{N}$, and every number of arms $k \in \mathbb{N}_{\geqslant 2}$, it holds that:*

$$\sup_{\substack{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^k \\ \boldsymbol{\nu} \text{ fulfills Assumption 6}}} R_{\boldsymbol{\nu},T}(\pi) \geqslant c_3 k^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}}, \tag{4.8}$$

where $c_3 > 0$ is a constant independent of u, ϵ, k and T.

Since (i) achieving the regret of Theorem 24 without further assumptions is not possible (Theorems 25 26) and (ii) Assumption 6 does not change the complexity of the non-adaptive case (Theorem 27), it makes sense to search for adaptive algorithms matching Theorem 24 under Assumption 6.

4.1.4. Trimmed Mean Estimator with Empirical Threshold

In this section, we present our novel *trimmed mean with empirical threshold* estimator, in which the threshold is computed from data. The trimmed mean estimator (Bickel, 1965), common in heavy-tailed statistics, cuts off the observations outside a predefined interval $[-M, M]$ with $M \geqslant 0$, named *trimming threshold*. Given a set of $s \in \mathbb{N}_{\geqslant 1}$ i.i.d. random variables $\mathbf{X} = \{X_1, \dots, X_s\}$, with expected value $\mu := \mathbb{E}[X_1]$, the trimmed mean estimator with threshold M is defined as:

$$\widehat{\mu}_s(\mathbf{X}; M) := \frac{1}{s} \sum_{j \in [s]} X_j \mathbb{1}_{\{|X_j| \leqslant M\}}. \tag{4.9}$$

The following result shows that, under truncated non-positivity (Assumption 6), it is possible to design an upper confidence bound on μ based on the trimmed mean estimator $\widehat{\mu}_s(\mathbf{X}; M)$ that can be computed with no knowledge of ϵ and u, depending only on the trimming threshold M.

Lemma 28 ((ϵ, u)-free Upper Confidence Bound)**.** *Let $\delta \in (0, 1/2)$ and $\mathbf{X} = \{X_1, \dots, X_s\}$ be a*

set of $s \in \mathbb{N}_{\geqslant 2}$ i.i.d. random variables satisfying $X_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, $\mu := \mathbb{E}[X_1]$, and $M > 0$ be a (possibly random) trimming threshold independent of $\mathbf{X}$. Then, under Assumption 6, it holds that:

$$\mathbb{P}\Big(\mu - \widehat{\mu}_s(\mathbf{X}; M) \leqslant \sqrt{\frac{2V_s(\mathbf{X}; M)\ln\delta^{-1}}{s}} + \frac{10M\ln\delta^{-1}}{s}\Big) \geqslant 1 - 2\delta, \tag{4.10}$$

where $V_s(\mathbf{X}; M)$ is the sample variance of the trimmed random variables, defined as:

$$V_s(\mathbf{X}; M) := \frac{1}{s-1}\sum_{j\in[s]}(X_j \mathbb{1}_{\{|X_j|\leqslant M\}} - \widehat{\mu}_s(\mathbf{X}; M))^2. \tag{4.11}$$

The result is obtained by applying the *empirical Bernstein's inequality* (Maurer and Pontil, 2009) and it is a *one-sided inequality* because of the nature of Assumption 6. From an algorithmic perspective, this enables us to build an optimistic index that does not require knowing the values of ϵ and u and represents the essential role of Assumption 6 in our `AdaR-UCB` algorithm.

The next step consists of computing the trimming threshold M in a fully data-driven way. Notice that the trimming threshold in `Robust UCB` (Bubeck et al., 2013a) is selected thanks to the knowledge of ϵ and u as $\widetilde{M}_s(\delta) = \left(\frac{us}{\ln\delta^{-1}}\right)^{\frac{1}{1+\epsilon}}$. Instead, we follow a procedure similar to that of Wang et al. (2021) for Huber regression, and we estimate an *empirical trimming threshold* via a root-finding problem. Specifically, given a set of $s \in \mathbb{N}_{\geqslant 1}$ i.i.d. random variables $\mathbf{X}' = \{X'_1, \dots, X'_s\}$ (independent of $\mathbf{X}$), the empirical threshold $\widehat{M}_s(\delta)$ is the solution of the equation:[6]

$$f_s(\mathbf{X}'; M, \delta) := \frac{1}{s}\sum_{j\in[s]}\frac{\min\{(X'_j)^2, M^2\}}{M^2} - \frac{c\ln\delta^{-1}}{s} = 0, \tag{4.12}$$

where $c > 0$ is a hyperparameter that will be set later. If the number of non-zero samples X'_j is sufficiently large, *i.e.*, $\sum_{j\in[s]}\mathbb{1}_{\{X'_j\neq 0\}} > c\ln\delta^{-1}$ (see Proposition 59 for details), Equation (4.12) admits a unique positive solution, that we denote as $\widehat{M}_s(\delta)$.[7] A reader might notice that we are solving the "sample version" of the "population version" equation $\mathbb{E}[f_s(\mathbf{X}'; M, \delta)] = 0$. Denoting with $M_s(\delta)$ the solution (when it exists) of this latter equation, we can establish a meaningful relation between $M_s(\delta)$ and the threshold $\widetilde{M}_s(\delta)$ used by `Robust UCB` (Bubeck

[6]The sets $\mathbf{X}$ and $\mathbf{X}'$ are chosen to have the same cardinality s to provide more readable results.

[7]An efficient algorithm for solving Equation (4.12) and its computational complexity analysis are reported in Appendix D.3.

et al., 2013a):

$$c \ln \delta^{-1} = \mathbb{E}[\min\{(X'_1)^2/M_s(\delta)^2, 1\}] \leqslant \mathbb{E}[|X'_1|^{1+\epsilon}] M_s(\delta)^{-1-\epsilon} \tag{4.13}$$

$$\implies M_s(\delta) \leqslant \left(\frac{us}{c \ln \delta^{-1}}\right)^{\frac{1}{1+\epsilon}} = c^{-\frac{1}{1+\epsilon}} \widetilde{M}_s(\delta). \tag{4.14}$$

In practice, however, we cannot solve the "population" equation $\mathbb{E}[f_s(\mathbf{X}'; M, \delta)] = 0$ and we need to resort to the "sample version", delivering $\widehat{M}_s(\delta)$. The following result shows that $\widehat{M}_s(\delta)$ behaves (in high probability) analogously to $M_s(\delta)$, for a suitable choice of c.

Theorem 29 (Bounds on $\widehat{M}_s(\delta)$)**.** *Let $\delta \in (0, 1/2)$ and $\mathbf{X}' = \{X'_1, \dots, X'_s\}$ be a set of $s \in \mathbb{N}_{\geqslant 1}$ i.i.d. random variables satisfying $X'_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, and let $\widehat{M}_s(\delta)$ be the (random) positive root of Equation* (4.12) *with $c > 2$. Then, if $\widehat{M}_s(\delta)$ exists, with probability at least $1 - 2\delta$, it holds that:*

$$\widehat{M}_s(\delta) \leqslant \left(\frac{us}{(\sqrt{c} - \sqrt{2})^2 \ln \delta^{-1}}\right)^{\frac{1}{1+\epsilon}} \quad \textit{and} \quad \mathbb{P}\left(|X_1| > \widehat{M}_s(\delta)\right) \leqslant (\sqrt{c} + \sqrt{2})^2 \frac{\ln \delta^{-1}}{s}. \tag{4.15}$$

The proof relies on the concentration inequalities for *self-bounding functions* (Maurer, 2006; Maurer and Pontil, 2009). By selecting $c > (1 + \sqrt{2})^2$, we have that, with probability $1 - 2\delta$, the empirical threshold $\widehat{M}_s(\delta)$ is smaller than $\widetilde{M}_s(\delta)$, used in `Robust UCB`. Furthermore, in Bubeck et al. (2013a), the particular form of the *deterministic* threshold $\widetilde{M}_s(\delta)$ allows the authors to apply *Bernstein's inequality* and obtain a concentration bound explicitly depending on ϵ and u (Lemma 1):

$$\mathbb{P}\left(\left|\widehat{\mu}_s(\mathbf{X}; \widetilde{M}_s(\delta)) - \mu\right| \leqslant 4u^{\frac{1}{1+\epsilon}} \left(\frac{\ln \delta^{-1}}{s}\right)^{\frac{\epsilon}{1+\epsilon}}\right) \geqslant 1 - 2\delta, \tag{4.16}$$

We now show that using the *random* threshold $\widehat{M}_s(\delta)$, instead, still allows achieving analogous guarantees with just a slightly larger constant.

Theorem 30 ((ϵ, u)-dependent Concentration Bound)**.** *Let $\delta \in (0, 1/4)$, $\mathbf{X} = \{X_1, \dots, X_{s/2}\}$, and $\mathbf{X}' = \{X'_1, \dots, X'_{s/2}\}$ be two independent sets of $s/2 \in \mathbb{N}_{\geqslant 2}$ i.i.d. random variables satisfying $X_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, $\mu := \mathbb{E}[X_1]$, and let $\widehat{M}_s(\delta)$ be the (random) positive root of Equation* (4.12) *with $c = (1 + \sqrt{2})^2$. Then, if $\widehat{M}_s(\delta)$ exists, it holds that:*

$$\mathbb{P}\left(\left|\widehat{\mu}_s(\mathbf{X}; \widehat{M}_s(\delta)) - \mu\right| \leqslant 8u^{\frac{1}{1+\epsilon}} \left(\frac{\ln \delta^{-1}}{s}\right)^{\frac{\epsilon}{1+\epsilon}}\right) \geqslant 1 - 4\delta. \tag{4.17}$$

First, we notice that, at the price of a slightly larger constant, this concentration bound displays the same behavior as that of Equation (4.16). However, remarkably, our estimator is fully

data-driven as the trimming threshold $\widehat{M}_s(\delta)$ is computed from dataset $\mathbf{X}'$ (*i.e.,* half of the available samples) with no knowledge of ϵ and u. Second, differently from Lemma 28, this is a *double-sided inequality* and holds even without Assumption 6. From a technical perspective, this result is proved by resorting to *Bernstein's inequality* and Theorem 29 to control the values of the estimated threshold $\widehat{M}_s(\delta)$.

Summarizing, we conclude that the trimmed mean estimator $\widehat{\mu}_s(\mathbf{X}; \widehat{M}_s(\delta))$ with the empirical threshold $\widehat{M}_s(\delta)$ fulfills two important properties: (i) under Assumption 6, it enjoys an upper confidence bound that is fully empirical and (ϵ, u)-free (Lemma 28). This bound will be used in the implementation of the `AdaR-UCB` algorithm; (ii) it enjoys (up to constants) the same concentration properties as the truncated mean with the (ϵ, u)-dependent threshold $\widetilde{M}_s(\delta)$ (Theorem 30). This bound, instead, will be used in the analysis of the `AdaR-UCB` algorithm.

4.1.5. An (ϵ, u)-Adaptive Approach for Heavy-Tailed Bandits

In this section, we address Research Question 2 by presenting `Adaptive Robust UCB` (`AdaR-UCB`, Algorithm 10), an (ϵ, u)-adaptive *anytime* regret minimization algorithm able to operate in the heavy-tailed bandit problem *with no prior knowledge on* ϵ *or* u, and providing its regret analysis.

The `AdaR-UCB` Algorithm

`AdaR-UCB` (Algorithm 10) is based on the optimism in-the-face-of-uncertainty principle, and built upon the `Robust UCB` strategy from Bubeck et al. (2013a) leveraging the estimator presented in Section 4.1.4. `AdaR-UCB` keeps track of the number of times every arm $i \in [k]$ has been selected $N_i(\tau)$ and maintains two disjoint sets of rewards $\mathbf{X}_i(\tau)$ and $\mathbf{X}'_i(\tau)$ (line 1). Specifically, $\mathbf{X}'_i(\tau)$ will be employed to compute the empirical threshold, while $\mathbf{X}_i(\tau)$ for the trimmed mean estimator. The algorithm operates over $\lfloor T/2 \rfloor$ rounds, indexed by τ, and, in every round $\tau \in \lfloor T/2 \rfloor$, it collects *two* samples from the selected arm I_τ (the time index is $t = 2\tau$). Specifically, `AdaR-UCB` first computes the *upper confidence bound* index $B_i(\tau)$ for every arm $i \in [k]$. If the condition for the existence of the positive root of Equation (4.12) is not verified (line 4), the index $B_i(\tau)$ is set to $+\infty$ (line 5), forcing the algorithm to pull arm i. Instead, if the condition is verified, the empirical threshold is computed $\widehat{M}_i(\tau) \leftarrow \widehat{M}_{N_i(\tau-1)}(\tau^{-3})$ (line 7) according to Equation (4.12) with $c = (1 + \sqrt{2})^2$ using the dataset $\mathbf{X}'_i(\tau - 1)$ and selecting $\delta = \tau^{-3}$. Then, the algorithm employs it (line 8) to compute the trimmed mean estimator $\widehat{\mu}_i(\tau) \leftarrow \widehat{\mu}_{N_i(\tau-1)}(\mathbf{X}_i(\tau - 1); \widehat{M}_i(\tau))$ (as in Equation 4.9) and variance estimator $V_i(\tau) \leftarrow V_{N_i(\tau-1)}(\mathbf{X}_i(\tau - 1); \widehat{M}_i(\tau))$ (as in Equation 4.11) using the samples from the other dataset $\mathbf{X}_i(\tau - 1)$. These quantities are then employed for the optimistic index computation

Algorithm 10: `Adaptive Robust UCB` (`AdaR-UCB`).

1 Initialize counters $N_i(0) = 0$, reward sets $\mathbf{X}_i(0) = \{\}$, $\mathbf{X}'_i(0) = \{\}$ for every $i \in [K]$, $\tau \leftarrow 1$, $t \leftarrow 2\tau$
2 **while** $\tau \leqslant \lfloor T/2 \rfloor$ **do**
3 **for** $i \in [K]$ **do**
4 **if** $N_i(\tau - 1) = 0$ **or** $\sum_{X' \in \mathbf{X}'_i(\tau-1)} \mathbb{1}_{\{X' \neq 0\}} \leqslant 4 \ln \tau^{-3}$ **then**
5 Compute the optimistic index: $B_i(\tau) = +\infty$
6 **else**
7 Compute the trimming threshold: $\widehat{M}_i(\tau) \leftarrow \widehat{M}_{N_i(\tau-1)}(\tau^{-3})$ solving the equation $f(\mathbf{X}'_i(\tau-1); M, \tau^{-3}) = 0$ (Eq. 4.12 with $c = (1+\sqrt{2})^2$)
8 Compute the trimmed mean estimator: $\widehat{\mu}_i(\tau) \leftarrow \widehat{\mu}_{N_i(\tau-1)}(\mathbf{X}_i(\tau-1); \widehat{M}_i(\tau))$ (Eq. 4.9) and the variance estimator $V_i(\tau) \leftarrow V_{N_i(\tau-1)}(\mathbf{X}_i(\tau-1); \widehat{M}_i(\tau))$ (Eq. 4.11)
9 Compute the optimistic index:

$$B_i(\tau) = \widehat{\mu}_i(\tau) + \sqrt{\frac{2V_i(\tau)\ln\tau^3}{N_i(\tau-1)}} + \frac{10\widehat{M}_i(\tau)\ln\tau^3}{N_i(\tau-1)}$$

10 **end**
11 **end**
12 Select arm $I_\tau \in \arg\max_{i\in[K]} B_i(\tau)$, play it **twice**, and receive rewards X and X'
13 Update reward sets $\mathbf{X}_{I_\tau}(\tau) = \mathbf{X}_{I_\tau}(\tau-1) \cup \{X\}$, $\mathbf{X}_i(\tau) = \mathbf{X}_i(\tau-1)$ for every $i \neq I_\tau$
14 Update reward sets $\mathbf{X}'_{I_\tau}(\tau) = \mathbf{X}'_{I_\tau}(\tau-1) \cup \{X'\}$, $\mathbf{X}'_i(\tau) = \mathbf{X}'_i(\tau-1)$ for every $i \neq I_\tau$
15 Update counters $N_i(\tau) = |\mathbf{X}_i(\tau)|$ for every $i \in [K]$, $\tau \leftarrow \tau + 1$, $t \leftarrow 2\tau$
16 **end**

$B_i(\tau)$ (line 9) according to the empirical bound of Lemma 28. The optimistic arm I_τ is then played *twice* (line 12) and the two collected samples are used to augment the reward sets $\mathbf{X}_i(\tau)$ and $\mathbf{X}'_i(\tau)$, respectively (lines 13-14), and the arm pull counters $N_i(\tau)$ (line 15).

Regret Analysis

In this section, we provide the regret analysis of `AdaR-UCB` under the truncated non-positivity assumption (Assumption 6). Let $\pi^{\texttt{AdaR-UCB}}$ be the policy defined by `AdaR-UCB`. We start with the instance-dependent regret bound.

Theorem 31 (Instance-Dependent Regret bound of `AdaR-UCB`)**.** *Let $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^k$ and $T \in \mathbb{N}_{\geqslant 2}$ be the learning horizon. Under Assumption 6, `AdaR-UCB` suffers a regret bounded as:*

$$R_{\boldsymbol{\nu},T}(\pi^{\texttt{AdaR-UCB}}) \leqslant \sum_{i:\Delta_i>0} \left[\left(120 \left(\frac{u}{\Delta_i} \right)^{\frac{1}{\epsilon}} + \frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)} \right) \ln \frac{T}{2} + 20\Delta_i \right]. \tag{4.18}$$

Some observations are in order. We notice that the dependence on ϵ and u match the instance-

dependent lower bound for the non-adaptive case (Bubeck et al., 2013a). Note that the trimming threshold estimation requires in AdaR-UCB the *forced exploration* (line 5) and leads to the additional logarithmic term $\sum_{i:\Delta_i>0} \frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)} \ln \frac{T}{2}$ that grows proportionally to the suboptimality gap Δ_i and inversely with the probability $\mathbb{P}_{\nu_i}(X \neq 0)$ of sampling a non-zero reward. This is explained by the condition (line 4) for the existence of a positive trimming threshold that requires a sufficiently large number of non-zero rewards. It is worth noting that for *absolutely continuous* reward distributions, *i.e.,* the ones we are interested in the heavy-tail setting, we have $\mathbb{P}_{\nu_i}(X \neq 0) = 1$. Moreover, if there is an arm i s.t. $\mathbb{P}_{\nu_i}(X \neq 0) = 0$, then based on Assumption 6, this arm is considered optimal. Consequently, AdaR-UCB achieves low regret by repeatedly selecting this arm. In such a case, this additional regret term reduces to $24 \sum_{i:\Delta_i>0} \Delta_i \ln \frac{T}{2}$, a term that was present in the regret bound of Robust UCB with the Catoni estimator too.[8] In general, we are unsure whether this term is unavoidable or an artifact of our algorithm and/or analysis. From a technical perspective, the proof of Theorem 31 follows similar steps to the result provided by Bubeck et al. (2013a) concerning the upper bound on regret for Robust UCB, although additional care is needed to control simultaneously the concentration of the empirical threshold and of the trimmed mean estimator. In conclusion, this result positively answers our Research Question 2, showing how AdaR-UCB nearly matches the instance-dependent lower bound for the non-adaptive case.

Finally, to complement the analysis, we provide the worst-case regret bound for AdaR-UCB.

Theorem 32 (Worst-Case Regret bound of AdaR-UCB)**.** *Let $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^k$ and $T \in \mathbb{N}_{\geqslant 2}$ be the learning horizon. Under Assumption 6, AdaR-UCB suffers a regret bounded as:*

$$R_{\boldsymbol{\nu},T}(\pi^{\texttt{AdaR-UCB}}) \leqslant 46\left(k \ln \frac{T}{2}\right)^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i>0} \left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)} \ln \frac{T}{2} + 20\Delta_i\right).$$

This result matches the lower bound from Bubeck et al. (2013a), up to logarithmic terms.

4.1.6. Open Problems

In this paper, we studied the (ϵ, u)-*adaptive* heavy-tailed bandit problem, where no information on moments of the reward distribution, not even which of them are finite, is provided to the learner. Focusing on two appealing research questions, we have: (i) shown that, with no further assumptions, no adaptive algorithm can achieve the same worst-case regret guarantees as in the non-adaptive case; (ii) devised a novel algorithm (AdaR-UCB), based on a fully data-driven estimator, enjoying nearly optimal instance-dependent and worst-case regret, under the truncated

[8] We remark that this second term, although logarithmic, does not depend on the reciprocal of the suboptimality gaps and it is negligible when $1/\Delta_i \gg 1$ compared to the first one.

non-positivity assumption.

Future directions include: (i) investigating the role of the truncated non-positivity assumption, especially, whether weaker assumptions can be formulated; (ii) characterizing the *limits of ϵ-adaptivity*, *i.e.*, the best performance attainable by an ϵ-adaptive algorithm *without additional assumption*; (iii) understanding whether the forced exploration for the empirical threshold computation in `AdaR-UCB` (and the corresponding regret term) is unavoidable.

4.2. Regret Mimimization in Piecewise-Stationary Heavy-Tailed Bandits

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 section, we tackle the heavy-tailed piecewise-stationary bandit problem. 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.

This section presents Genalti et al. (2025), a joint project with Sujay Bhatt, Nicola Gatti and Alberto Maria Metelli, currently under review in a specialistic venue.

4.2.1. Introduction

In this section, we focus on a broad class of problems that relaxes, at the same time, two core assumptions of the standard MAB problem. In particular, we focus on *heavy-tailed non-stationary* MABs. As in the previous section, 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. Moreover, it extends the assumption of sub-gaussian 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 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 over 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. The contributions are organized as follows:

- In Section 4.2.2, we introduce the definition of the heavy-tailed piecewise-stationary MAB setting. We define the learning problem and introduce a lower bound on the expected regret for this setting.
- In Section 4.2.3, we recall some notions and results from the existing literature on mean estimation for heavy-tailed random variables and on change-point detection.
- In Section 4.2.4, we introduce `Robust-CPD-UCB`, an algorithm from regret minimization in our setting. We provide theoretical guarantees on its expected regret and insights on choosing its parameters.
- Finally, in Section 4.2.5 (and Appendix E.3), we provide numerical evaluations of the performance of `R-CPD-UCB`, comparing it with baselines from the literature on both real-world and synthetic data.

4.2.2. 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 for the problem.

Bandit Settings

We formally introduce the heavy-tailed bandit setting again, but in a slightly different way from the previous section, as we now want to deal with the *centered* $(1+\epsilon)$-th order moment. The difference is not much, as one can be trivially upper bounded by the other, but it becomes relevant here to perform more natural calculations.

Heavy-Tailed Bandits. In *heavy-tailed* bandits Bubeck et al. (2013a), the probability distributions $\{\nu_i\}_{i=1}^k$ are *heavy-tailed*. In this work, we use the same definition of heavy-tailed MAB (HT MAB, for short).

Definition 4.1 (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*

$$\mathbb{E}_\nu[|X - \mathbb{E}_\nu[X]|^{1+\epsilon}] \leqslant v, \tag{4.19}$$

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 (4.19)*, we call $\boldsymbol{\nu}$ a heavy-tailed bandit (HT MAB, for short).*

Note that Equation (4.19) implies that the variance of the rewards-generating distributions may be infinite (when $\epsilon < 1$). Most of the technical tools employed for sub-gaussian 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) 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 4.2 (Piecewise-Stationary Bandit)**.** *Let $\{\boldsymbol{\nu}^{(j)}\}_{j\in[\Upsilon]}$ be a set of Υ MABs. Then, let Υ be a set of timesteps $\{t_c^{(j)}\}_{j\in[\Upsilon]} \subset [T]$ 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 the j-th epoch of the PS MAB, and $t_c^{(j)}$ to the j-th breakpoint. To simplify notation, we define $\mu_i^{(j)} := \mathbb{E}_{\nu_i^{(j)}}[X_{i,t}]$ 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 $\delta_i^{(j)} := |\mu_i^{(j)} - \mu_i^{(j-1)}|$ 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, $E_0 = \varnothing$ and $\delta_i^{(0)} = \infty$ for every $i \in [k]$. In Appendix E.2, 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 (4.19). We call this setting the *heavy-tailed piecewise-stationary* setting (HTPS, for short).

Definition 4.3 (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* $\mathcal{B}_{(v,\epsilon,\Upsilon,\boldsymbol{t}_c)}$*, where* $\boldsymbol{t}_c = \{t_c^{(j)}\}_{j\in[\Upsilon]}$.

Definition 4.3 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.

Learning Goal

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* $R_T(\pi)$[9], *i.e.*, the cumulative performance gap w.r.t. to the best *policy* over a learning horizon.

Definition 4.4 (Expected Cumulative Regret)**.** *Given a policy* π*, we define the expected cumulative regret of* π *as:*

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

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

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.*, $i_j^* \in \arg\max_{i\in[k]} \mu_i^{(j)}$. 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.

[9]Differently from Chapter 2, we omitted the instance parameters from the pedices for the sake of the clarity.

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

$$\sup_{\nu \in \mathcal{B}_{(v,\epsilon,\Upsilon)}} R_T(\pi) \geqslant \frac{1}{25}(k\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}. \tag{4.20}$$

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., 2013a), while in PS MABs the lower bound is $\Omega(\sqrt{k\Upsilon T})$ (Garivier and Moulines, 2011). Thus, Equation (4.20) is a natural combination of these two results that can be recovered by either setting $\Upsilon = 1$ or $\epsilon = 1$. We refer to Appendix E.1 for the proof.

4.2.3. 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*.

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 proved to achieve a sub-optimal concentration (Bubeck et al., 2013a). However, alternative estimators enjoying optimal rates have been proposed. We focus on the elegant *Catoni estimator* (Catoni, 2012), defined using a *Catoni-type influence function* $\phi : \mathbb{R} \to \mathbb{R}$. The Catoni estimator $\widehat{\mu}_c$ for a sequence of variables $\{X_i\}_{i=1}^n$ is the solution of:

$$\sum_{i=1}^{n} \phi_\epsilon(\lambda_i(X_i - \widehat{\mu}_c)) = 0, \quad \text{where} \quad \phi_\epsilon(x) = \log\left(1 + |x| + \frac{|x|^{1+\epsilon}}{1+\epsilon}\right), \tag{4.21}$$

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. (2022a).

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 sub-gaussian 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:

$$\mathbb{P}(\forall t \in \mathbb{N}^+ : \mu \in \text{CI}_t) \geqslant 1 - \gamma. \tag{4.22}$$

The random intervals $\{\text{CI}_t\}_{t\in\mathbb{N}^+}$ that satisfy property (4.22) 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 $w(t, P, \gamma) := \sup_{\mu_1,\mu_2\in CI_t} |\mu_1 - \mu_2| \leqslant w(t, P, \gamma)$, 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 $t_c \geqslant 1$ be an unknown breakpoint, *i.e.*, $P_t = P_0$ for every $t \leqslant t_c$ and $P_t = P_1$ 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, 2023b,a) shed 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 (2023a). `repeated-FCS-detector` is a meta-algorithm that requires a CS computation strategy and uses it as a black-box tool, defined as:

Definition 4.6 (`repeated-FCS-detector`, Shekhar and Ramdas (2023a)). *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 $CS^{(t)} := \{CI_n^{(t)}\}_{n\geqslant t}$, formed using samples X_t, X_{t+1}, X_{t+2}, and onwards. Moreover, we update all previously initialized CS $\{CI^{(i)}\}_{i<t}$ 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., $\tau = \inf_{t\in\mathbb{N}}\{\bigcap_{n=0}^{t} CS^{(n)} = \varnothing\}$.

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 (2023a), theoretical guarantees on both the ARL and the detection delay of `repeated-FCS-detector` are provided expressed in terms of the width of the $(1-\gamma)$-CS provided to the detector. We now report the theoretical guarantees of `repeated-FCS-detector`.

Theorem 4.7 (Guarantees of `repeated-FCS-detector`, Shekhar and Ramdas (2023a)). *Consider a CPD problem with observations $\{X_t\}_{t\in\mathbb{N}}$ i.i.d. from $P_0 \in \mathcal{P}^{\mu_0}$ for $t \leqslant t_c$ and from $P_1 \in \mathcal{P}^{\mu_1}$ for $t > t_c$. Let $\delta := |\mu_1 - \mu_0|$. Suppose we can construct $(1-\gamma)$-confidence sequences with pointwise width $w(t, P_0, \gamma)$ and $w(t, P_1, \gamma)$ for pre- and post-change mean, respectively. Then, we have: **(i)** When there is no changepoint, the `repeated-FCS-detector` satisfies $\mathbb{E}_\infty[\tau] \geqslant \frac{1}{\gamma}$; **(ii)** Suppose $t_c < \infty$ and large enough to ensure that $w(t_c, \mu_0, \gamma) < \delta$. Introduce the event $\mathcal{E} = \{\mu_0 \in \bigcap_{t=1}^{t_c} CI_t^{(1)}\}$, and note that $\mathbb{P}(\mathcal{E}) \geqslant 1-\gamma$ by construction. Then, for $\gamma \in (0, 0.5)$, we have $\mathbb{E}_{t_c}[(\tau - t_c)^+ | \mathcal{E}] \leqslant \frac{3u_0(\mu_0,\mu_1,t_c)}{1-\gamma}$, where $u_0(\mu_0, \mu_1, t_c) := \min_{n\in\mathbb{N}}\{w(n, \mu_1, \gamma) + w(t_c, \mu_0, \gamma) < \delta\}$.*

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.

4.2.4. 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 this tool 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.

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

$$CI_t^\phi = \left\{ m \in \mathbb{R} : \sum_{i=1}^{t} \phi_\epsilon(\lambda_i(X_i - m)) \in \left[\mp \frac{v}{2} \sum_{i=1}^{t} \lambda_i^{1+\epsilon} \pm \log\left(\frac{2}{\gamma}\right) \right] \right\}, \tag{4.23}$$

where ϕ_ϵ *is the Catoni-type influence function (Equation 4.21).*

From now on, we call *Catoni CS* the confidence sequence defined as in Equation (4.23). 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 E.1) that Equation (4.23) represents a proper $(1-\gamma)$-CS for the mean, attaining an optimal width (Bhatt et al., 2022a). 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 4.1 (Detection Delay of Catoni-FCS-detector)**.** *Consider a CPD problem with observations* $\{X_t\}_{t\in\mathbb{N}}$ *drawn i.i.d. from* $P_0 \in \mathcal{H}_{\epsilon,v} \cap \mathcal{P}^{\mu_0}$ *for* $t \leqslant t_c$ *and from* $P_1 \in \mathcal{H}_{\epsilon,v} \cap \mathcal{P}^{\mu_1}$ *for* $t > t_c$*. Let* $\delta := |\mu_1 - \mu_0|$*. Suppose that there exists a known upper bound* T *of the change point (*$t_c \leqslant T$*). Let* $n_{min} := 68\log(T^{\frac{1+\epsilon}{\epsilon}})$ *and suppose* $t_c \geqslant n_{min}$ *large enough s.t.* $w(t_c, P_0, \gamma) \leqslant \frac{\delta}{2}$*. Set* $\gamma = \frac{2}{T^3}$*. Then, there exists a predictable sequence* $\{\lambda_i\}_{i=1}^T$ *s.t.* Catoni-FCS-detector *enjoys* ***(i)*** $\mathbb{P}_{t_c}\left((\tau - t_c)^+ \leqslant \mathcal{O}\left(v^{\frac{1}{\epsilon}} \frac{\log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}}\right)\right) \geqslant 1 - \frac{14}{T}$ *and* ***(ii)*** $\mathbb{P}_{t_c}(\tau < t_c) \leqslant \frac{14}{T}$.

We point out the importance of this specialized result. Since the guarantees of Theorem 4.7 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 4.7 explicit by using the properties of Catoni CS. Due to space reasons, the proof is postponed to Appendix E.1. 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),

Algorithm 11: `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* $\lfloor k/\eta \rfloor \leqslant k$ **then**
    Select and play $I_t \leftarrow t' \mod \lfloor k/\eta \rfloor$.
  else
    Update π_s with the history of the last t' rounds.
    Select and play I_t according to π_s.
  Receive X_t and update $N_{I_t,t} \leftarrow N_{I_t,t} + 1$ and $t' \leftarrow t' + 1$.
  if $N_{I_t,t} \geqslant n_{min}$ **then**
    Start a new $(1-\gamma)$-CS $\mathrm{CS}_{I_t}^{(t')}$ for action I_t, according to Equation (4.23).
    if $\exists a, b \in [t'] : CS_{I_t}^{(a)} \cap CS_{I_t}^{(b)} = \varnothing$ **then**
      Reset $t' \leftarrow 0$, $N_{i,t} \leftarrow 0 \quad \forall i \in [k]$.
      Remove all initialized CS.

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.

Remark 9 (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 10 (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.*

`Robust-CPD-UCB`

In this section, we introduce `Robust-CPD-UCB` (`R-CPD-UCB` for short, Algorithm 11), an algorithm for PS HT bandits. `R-CPD-UCB` actively adapts to the changes in the reward-generating distribution. The algorithm has three components: (1) a sub-algorithm suited for

the stationary 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 11 proceeds as follows: roughly every $\lfloor k/\eta \rfloor$ rounds it tries all the actions once (lines 4-5), this 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 point occurred (lines 12-14), in this case, everything is reset (line 15-16).

Remark 11. *(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). Moreover, as most of the algorithms for the PS setting, ours share the usage of a stationary bandit sub-routine. However, a crucial difference relies on 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 hyper-parameters, 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 12. *(Priori Knowledge of* `R-CPD-UCB`*) Algorithm 11 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. Thus, in principle, our algorithm requires knowledge of T only. In practice, this property ensures that no tuning must happen.*

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

This assumption 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 Assumption 4.9 is a technical assumption aimed at the theoretical analysis of the algorithm. Algorithm 11 can operate regardless of this assumption, as shown in Section 3.1.6. We are now ready to present our main result.

Theorem 4.10 (Regret Upper Bound of R-CPD-UCB)**.** *Under Assumption 4.9,* R-CPD-UCB *suffers an expected cumulative regret bounded as:*

$$R_T(\pi^{\text{R-CPD-UCB}}) \leqslant \mathcal{O}\Bigg(\underbrace{\sum_{j=1}^{\Upsilon} \frac{v^{\frac{1}{\epsilon}} \log(T)}{(\widetilde{\delta}_{min}^{(j)})^{\frac{1+\epsilon}{\epsilon}}} \left\lceil \frac{k}{\eta} \right\rceil \Delta_{max}^{(j)}}_{\text{(A) Detection Delay Contribution}} + \underbrace{\sum_{j=1}^{\Upsilon} \mathbb{E}[R^{\pi_s}(|E_j|)]}_{\text{(B) Stationary Policy Regret}} + \underbrace{\eta \sum_{j=1}^{\Upsilon} |E_j| \Delta_{max}^{(j)}}_{\text{(C) Uniform Exploration}} \Bigg). \tag{4.24}$$

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 can leverage the methodology developed in Besson et al. (2022) and obtain the following result. 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:

$$\text{(A)} \leqslant \frac{v^{\frac{1}{\epsilon}} \sqrt{k \Upsilon T \log(T)}}{\eta_0 \delta_{min}^{\frac{1+\epsilon}{\epsilon}}} \Delta_{max}, \qquad \text{(C)} \leqslant \eta_0 \sqrt{k(\Upsilon+1) T \log(T)} \Delta_{max}. \tag{4.25}$$

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., 2013a, Section 2.2). As a result, we get the following bounds. Let π_s be the Robust UCB policy with median-of-means estimator (Bubeck et al., 2013a, Section 2.2). Under Assumption 4.9, R-CPD-UCB suffers an expected cumulative regret bounded as:

$$\mathbb{E}[R^{\pi^{R\text{-}CPD\text{-}UCB}}(T)] \leqslant \mathcal{O}\left((A) + \underbrace{\sum_{j=1}^{\Upsilon} \sum_{i:\Delta_i^{(j)}>0} \frac{v^{\frac{1}{\epsilon}} \log(|E_j|)}{(\Delta_i^{(j)})^{\frac{1}{\epsilon}}}}_{(B_1)\ \textit{Robust UCB Regret (Instance Dependent)}} + (C) \right). \tag{4.26}$$

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

$$R_T(\pi^{R\text{-}CPD\text{-}UCB}) \leqslant \widetilde{\mathcal{O}}((A) + \underbrace{(k\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}}_{(B_2)\ \textit{Robust UCB Regret (Instance Independent)}} + (C)). \tag{4.27}$$

Equation (4.26) is a direct consequence of Theorem 4.10 and Theorem 3 of Bubeck et al. (2013a). Equation (4.27) follows from Theorem 4.10, Proposition 1 of Bubeck et al. (2013a), 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. Let π_s be the Robust UCB policy with median-of-means estimator from (Bubeck et al., 2013a, 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 4.9, R-CPD-UCB using η_{j+1} after the j-th detection suffers an expected cumulative regret bounded as:

$$R_T(\pi^{R\text{-}CPD\text{-}UCB}) \leqslant \mathcal{O}\left(\frac{v^{\frac{1}{\epsilon}}\sqrt{k\Upsilon T\log(T)}}{\eta_0 \delta_{min}^{\frac{1+\epsilon}{\epsilon}}}\Delta_{max} + \frac{k\Upsilon v^{\frac{1}{\epsilon}}\log(T/\Upsilon)}{\Delta_{min}^{\frac{1}{\epsilon}}} \right). \tag{4.28}$$

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

$$R_T(\pi^{R\text{-}CPD\text{-}UCB}) \leqslant \widetilde{\mathcal{O}}((k\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}). \tag{4.29}$$

Equation (4.28), 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 (4.29), 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 (4.29) matches, up to constants, the lower bound presented in Theorem 4.5. Thus, if we focus on the dependence on T, Υ, v, and k the performance guarantees of `R-CPD-UCB` are nearly-optimal.

4.2.5. 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 E.3 for additional details and experimental campaigns.

Casting Real-World Data to HTPS MABs

We model a real-world scenario as an HTPS MAB and, then, we leverage a real dataset to generate an HTPS MAB instance on which `R-CPD-UCB` is tested. **Setting.** We consider the problem of profit maximization in financial trading. As pointed out by Panahi (2016), financial data exhibit heavy tails. A financial application of HT MABs is identifying the most profitable cryptocurrency among k options. In fact, at the start of each day, an investor would like to invest a share of their money in the cryptocurrency with the highest closing price. This very same application has been studied, for example, in Yu et al. (2018) and Lee and Lim (2022), both in the context of HT MABs. We use the same dataset (Kaggle link) employed in Lee and Lim (2022). In Figure 4.1, we report the closing prices of four selected currencies among the top ten by market capitalization, along with a piecewise-constant fit of the data that minimizes the squared error. We observe two things: first, the piecewise constant approximation is a better fit than any constant approximation (in the previous works on HT MABs, the reward from a given currency was always considered stationary); second, this approximation suffers a high error in certain segments where the stochastic fluctuations are really strong. Following the existing literature Panahi (2016), we fit the price distribution inside every segment with a Pareto distribution having its mean centered on the segment height, using $\epsilon < 1$ and $v = 3$. Thus, the profit maximization problem in cryptocurrency trading can be treated as an HTPS MAB. **Results.** Starting from the piecewise-constant fit of the prices, we can build an HTPS MAB environment on which we test `R-CPD-UCB`, together with `Sliding Window UCB` Garivier and Moulines (2011) and `MR-APE` Lee and Lim (2022), which was already tested on the same dataset when assuming stationarity. In Figure 4.2, we report the cumulative regrets obtained

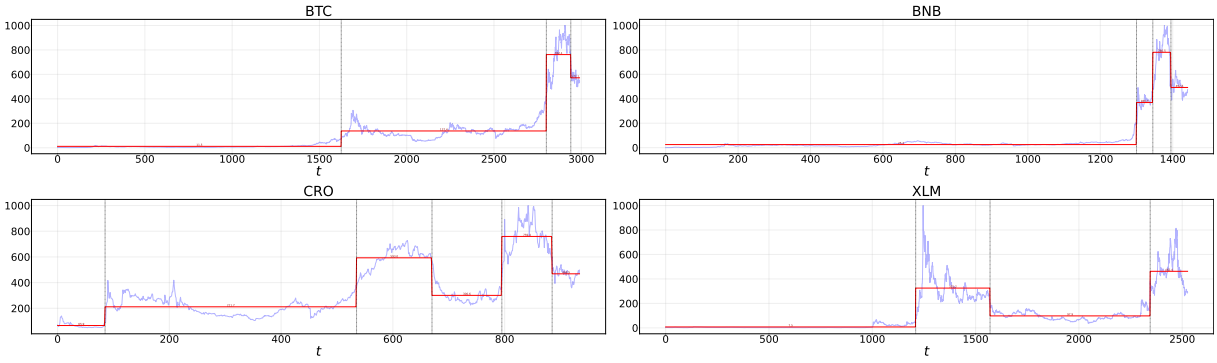


Figure 4.1: Rescaled closing prices of four selected cryptocurrencies (blue) with a piecewise-constant approximation (red). Each time step is a day starting in April 2016. Source: Kaggle Dataset.

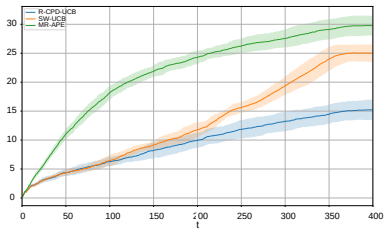


Figure 4.2: Cumulative regrets on HTPS built from cryptocurrency dataset. 20 trials, mean $\pm$ std.

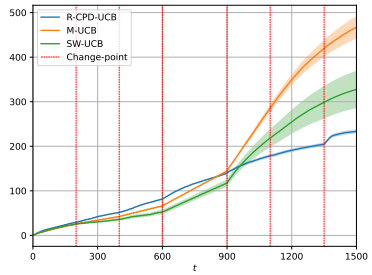


Figure 4.3: Gaussian rewards.

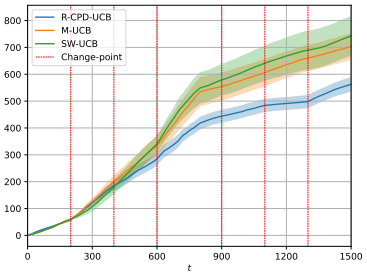


Figure 4.4: Pareto rewards.

Figure 4.5: Cumulative regrets. 20 trials, mean $\pm$ std.

by the three algorithms averaged over 20 trials (x-axis is rescaled). `R-CPD-UCB` performs better than the two competitors, as it is the only algorithm tackling both heavy-tailedness and non-stationarity of the setting.

Regret Minimization in Highly Non-Stationary Environments

We now evaluate how `R-CPD-UCB` behaves in highly dynamic scenarios where changes are close.

Setting. We compare `R-CPD-UCB` with two of the most popular algorithms from the literature, `Monitored UCB` 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 $\epsilon < \frac{1}{2}$. 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 4.9. **Results.** In Figure E.2, we report the cumulative regrets suffered by the considered algorithms. `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 E.2b). Interestingly, `R-CPD-UCB` can outperform both `Monitored UCB` and `Sliding Window UCB` even when the rewards are Gaussian. This is 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 4.9 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 4.9, in practice, is not very limiting.

4.2.6. Discussion and Future Directions

In this section, we provided the first study on regret minimization in heavy-tailed piecewise-stationary bandits. We provided a lower bound on the performance of every algorithm and proposed `Robust-CPD-UCB`, a novel algorithm whose regret nearly matches the lower bound. We leverage novel advancements in the theory of change-point detection, building `Catoni-FCS-detector`, a general detection strategy suited for distributions with infinite variance. Finally, numerical evaluation shows that the performance of `R-CPD-UCB` is solid when compared to existing baselines. An interesting future direction would be to study the HTPS MAB problem when v and ϵ are unknown.

5 | Conclusions and Future Directions

In this thesis, we presented online learning methods to deal with settings of practical and theoretical interest. For each of these settings, we presented algorithmic solutions, provided theoretical guarantees on their performance, and numerically validated them in both synthetic and real-world environments.

In Chapter 2, we provided a brief introduction to the multi-armed bandit problem and its generalization to Markov decision processes, providing key results from existing literature.

In Chapter 3, we studied problems at the intersection of MABs and MDPs. In general, MDPs cannot be tackled in the same way as MABs, and the possibilities for providing finite-time theoretical guarantees for online algorithms are limited. The novel settings introduced in Sections 3.1 and 3.2 are particular MDPs of interest, for which we provide MAB-style algorithms and offer strong finite-time guarantees on the expected cumulative regret. In Section 3.3, we tackled a well-established setting in the bandit literature, the rising bandit problem. We improved over existing results and provided novel technical tools that can be useful in neighboring settings.

In Chapter 4, we switched our interest to a generalization of the classic stochastic structure of the MAB problem. We relaxed the assumption for sub-Gaussian/bounded support reward distributions and studied heavy-tailed bandits. In Section 4.1 we tackled the open question of adaptivity to unknown distributional parameters. Our results shed light on the complexity of this problem, which was previously not studied in the literature. Heavy-tailed bandits are useful in modeling many real-world scenarios, such as finance and network routing. While being fully data-driven is helpful in real-world applications, the stationary rewards assumption is still too restrictive. In Section 4.2, we deal with the heavy-tailed piecewise-stationary bandit problem, that allows for the reward distributions to change their mean over time, in a piecewise-constant manner. We are the first to provide an algorithm to address this problem, and we provide its theoretical guarantees showing that they are order-optimal. Finally, we validate our approach using data from the real-world.

5.1. Future Directions and Open Problems

In what follows, we outline all of the interesting future research lines and the open questions that arise from this thesis.

5.1.1. Autoregressive Bandits

The ARB setting is general and flexible, but it is unclear whether some of the assumptions can be relaxed further. In particular, we draw attention to Assumption 1.a. We conjecture that this assumption can be relaxed to allow for negative coefficients, but at the cost of ensuring that they lie on a unitary disk An in-depth discussion about this can be found in Section A.3. In fact, our analysis is likely to work even under this weaker condition, possibly at the price of some slight modifications. However, the most interesting result in the ARB setting would be a lower bound for the regret. In particular, while we argue that most of the quantities appearing in the regret upper bound of `AutoRegressive Upper Confidence Bound` have a tight dependency, we still don't know if the dependence from m can be lowered from an order of $\frac{3}{2}$ to $\frac{1}{2}$. Also, there may be an extra $(1-\Gamma)$ term in the denominator. A tight lower bound would indicate whether these dependencies are truly necessary or not.

5.1.2. Graph-Triggered Bandits

If we focus on the rising GTB problem in particular, it is clear that the regret rate in the stochastic case can be lowered. While there is no lower bound for the GTB setting in particular, one can cast the bound to the simpler case of restless rising bandits, and observe that this is not tight (we actually proved this in Section 3.3). It would be interesting to plug the algorithm proposed in Section 3.3, which has a better order than the one from Metelli et al. (2022) on which our algorithms are based, into the GTB setting. This would probably lead to better bounds, at least in the dependency on T. Finally, an interesting direction is to search for other ways to let the rewards evolve for which the GTB setting can be addressed with strong theoretical guarantees. We focused on monotonically evolving bandits, *i.e.*, rising and rotting bandits, but recently the piecewise-stationary Yu and Mannor (2009) and the smooth Jia et al. (2023) bandits have emerged as interesting ways to model non-stationarity.

5.1.3. Restless Rising Bandits

In Section 3.3, we provided a lower bound on the regret for the rising concave setting in the order of $T^{\frac{3}{5}}$. Our algorithm, however, suffers a regret that is bounded in the order of $T^{\frac{7}{11}}$. While numerically close, there is still a gap between the two. We strongly conjecture that the true regret

suffered by our algorithm can be bounded with a tight order of $T^{\frac{3}{5}}$, but the analysis has to be refined. A future direction of research is thus to definitely close this gap.

5.1.4. Heavy-Tailed Bandits

In this thesis, Section 4.1 is the one providing the most open questions that would be interesting to address. As a matter of fact, an *Open Problem* paper (Genalti and Metelli, 2025) has been published at COLT 2025 to collect all of the interesting future directions to pursue on this topic.

We prove that it is not possible to achieve the same rate as when u and ϵ are known while adapting to them. Our lower bounds are, however, qualitative, and they do not quantify exactly what the additional cost is. Moreover, they address the adaptation to u and ϵ separately, and not simultaneously. A first problem to address is to quantify, in terms of extra regret paid, the cost of adaptation to u, ϵ, and simultaneously together. Then, one may draw its attention to providing positive results, *i.e.*, provide an adaptive algorithm that has tight regret rates w.r.t. to such lower bounds. Our algorithm, `AdaR-UCB`, achieves the same regret rates as the non-adaptive algorithms, but at the cost of Assumption 6. It would be interesting to understand if there exists a *minimal* assumption s.t. this is possible. We already know from Genalti and Metelli (2025) that our assumption can be relaxed and the tail bias can be bounded with the variance of the truncated distribution instead of 0. However, it is not clear if these kinds of assumptions are the best or if something weaker can be provided while maintaining a tight rate.

Of course, the problem of adaptation to the distributional parameters can be cast in the piecewise-stationary setting as well. In particular, it would be interesting to see if, with the addition of Assumption 6 or any other similar assumption, it is possible to still obtain the tight regret rates in the piecewise-stationary setting.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 24, 2011.

Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR, 2021.

Fahad Albalawi, Zihang Dong, and David Angeli. Regret-based robust economic model predictive control for nonlinear dissipative systems. In *European Control Conference (ECC)*, pages 1105–1111. IEEE, 2021.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pages 23–35. PMLR, 2015.

Kumar Ashutosh, Jayakrishnan Nair, Anmol Kagrecha, and Krishna Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR, 2021.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE annual foundations of computer science*, pages 322–331. IEEE, 1995.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

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.

Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, 22(14):1–38, 2021.

Francesco Bacchiocchi, Gianmarco Genalti, Davide Maran, Marco Mussi, Marcello Restelli, Nicola Gatti, and Alberto Maria Metelli. Autoregressive bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 937–945. PMLR, 2024.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *CoRR*, abs/1803.06971, 2018.

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.

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

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Nearly optimal catoni's m-estimator for infinite variance. In *International Conference on Machine Learning*, pages 1925–1944. PMLR, 2022b.

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.

Peter J Bickel. On some robust estimates of location. *The Annals of Mathematical Statistics*, pages 847–858, 1965.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL `https://doi.org/10.1093/acprof:oso/9780199535255.001.0001`.

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 21, 2008.

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

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR, 2013b.

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

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.

Leonardo Cella, Massimiliano Pontil, and Claudio Gentile. Best model identification: A rested bandit formulation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 1362–1372. PMLR, 2021.

Qinyi Chen, Negin Golrezaei, and Djallel Bouneffouf. Non-stationary bandits with auto-regressive temporal dependency. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 696–726. PMLR, 2019.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *JMLR Proceedings*, pages 208–214, 2011.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (2nd Edition)*. Wiley, 2006.

Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Annual Conference on Learning Theory (COLT)*, pages 355–366, 2008.

Ofer Dekel, Ambuj Tewari, and Raman Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the International Conference on Machine Learning (ICML)*. Omnipress, 2012a.

Ofer Dekel, Ambuj Tewari, and Raman Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012b.

J.L. Doob. *Stochastic Processes*. Probability and Statistics Series. Wiley, 1953.

Matteo Gagliolo and Jürgen Schmidhuber. Algorithm portfolio selection as a bandit problem with unbounded losses. *Annals of Mathematics and Artificial Intelligence*, 61:49–86, 2011.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

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

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.

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.

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.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto M. Metelli. Bridging rested and restless bandits with graph-triggering: Rising and rotting. *CoRR*, abs/2409.05980, 2024b.

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*, 2024c.

Gianmarco Genalti, Sujay Bhatt, Nicola Gatti, and Alberto Maria Metelli. Catoni-style change point detection for regret minimization in non-stationary heavy-tailed bandits, 2025. URL `https://arxiv.org/abs/2505.20051`.

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

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in k–armed bandits. *Journal of Machine Learning Research*, 24(13):1–33, 2023.

James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.

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.

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

Christine Herlihy and John P. Dickerson. Networked restless bandits with positive externalities. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11997–12004. AAAI Press, 2023.

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

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *international conference on machine learning*, pages 9173–9200. PMLR, 2022.

Prakirt Raj Jhunjhunwala, Sharayu Moharir, D Manjunath, and Aditya Gopalan. On a class of restless multi-armed bandits with deterministic policies. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 487–491. IEEE, 2018.

Su Jia, Qian Xie, Nathan Kallus, and Peter I. Frazier. Smooth non-stationary bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 14930–14944. PMLR, 2023.

Su Jia, Qian Xie, Nathan Kallus, and Peter I. Frazier. Smooth non-stationary bandits, 2024.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:15312–15325, 2020.

Richard M Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, pages 85–103, 1972.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 681–690. ACM, 2008.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, 3 edition, 2020.

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

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *IEEE Conference on Decision and Control (CDC)*, pages 757–762. IEEE, 2021.

Tor Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462, 2020.

OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3074–3083, 2017.

Yang Li, Jiawei Jiang, Jinyang Gao, Yingxia Shao, Ce Zhang, and Bin Cui. Efficient automatic cash via rising bandits. In *AAAI Conference on Artificial Intelligence*, 2020.

Jörg Liebeherr, Almut Burchard, and Florin Ciucu. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2): 1010–1024, 2012.

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.

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

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

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 23:32–1, 2022.

Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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

Cristiano Migali, Marco Mussi, Gianmarco Genalti, and Alberto Maria Metelli. Tightening regret lower and upper bounds in restless rising bandits. In *Advances in Neural Information Processing Systems*, 2025.

Marco Mussi, Alberto Maria Metelli, and Marcello Restelli. Dynamical linear bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 25563–25587. PMLR, 2023.

Marco Mussi, Alessandro Montenegro, Francesco Trovò, Marcello Restelli, and Alberto M. Metelli. Best arm identification for stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 36953–36989. PMLR, 2024a.

Marco Mussi, Alessandro Montenegro, Francesco Trovò, Marcello Restelli, and Alberto Maria Metelli. Best arm identification for stochastic rising bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2024b.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory (ALT)*, pages 214–228. Springer, 2012.

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.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with

delayed, aggregated anonymous feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4105–4113. PMLR, 2018.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvári, and Steffen Grünewälder. Bandits with delayed, aggregated anonymous feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 4102–4110. PMLR, 2018.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

Sartaj Sahni. Computationally related problems. *SIAM Journal on computing*, 3(4):262–279, 1974.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 2019.

Julien Seznec, Pierre Ménard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 3784–3794. PMLR, 2020.

Ron Shamir, Roded Sharan, and Dekel Tsur. Cluster graph modification problems. *Discrete Applied Mathematics*, 144(1-2):173–182, 2004.

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

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

Sho Takemori, Yuhei Umeda, and Aditya Gopalan. Model-based best arm identification for decreasing bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pages 1567–1575. PMLR, 2024.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

Francesco Trovò, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68: 311–364, 2020.

Jonas Umlauft and Sandra Hirche. Learning stable stochastic nonlinear dynamical systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3502–3510. PMLR, 2017.

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

Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021.

Siwei Wang, Longbo Huang, and John C. S. Lui. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 11878–11889. Curran Associates, Inc., 2020.

Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits. *IEEE Control Systems Letters*, 5(4):1423–1428, 2020.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The International Conference on Research & Development in Information Retrieval (SIGIR)*, pages 495–504. ACM, 2018.

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.

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.

A | Autoregressive Bandits

A.1. Proofs

Theorem 1 (Optimal Policy). *Under Assumption 1.a, for every round $t \in [T]$, the optimal policy $\pi_t^*(H_{t-1})$ satisfies:*

$$\pi_t^*(H_{t-1}) \in \arg\max_{i \in [k]} \langle \boldsymbol{\gamma}(i), \mathbf{Z}_{t-1} \rangle. \tag{3.5}$$

Proof We first prove an intermediate result auxiliary to get to the final statement. Let us denote with $J_T^*(\mathbf{Z})$ the expected cumulative reward when the initial observations vector is $\mathbf{Z} = (1, x_0, x_{-1}, \dots, x_{-k+1})$. Let us denote with $\geq$ the element-wise inequality. We show that for every $T \in \mathbb{N}$, if $\mathbf{Z} \geq \overline{\mathbf{Z}}$, then $J_T^*(\mathbf{Z}) \geqslant J_T^*(\overline{\mathbf{Z}})$.

We proceed by induction.

For $T = 1$, we have $J_1^*(\mathbf{Z}) = \max_{i \in [k]} \langle \boldsymbol{\gamma}(i), \mathbf{Z} \rangle = \langle \boldsymbol{\gamma}(i_1^*), \mathbf{Z} \rangle$, where $i_1^* \in \arg\max_{i \in [k]} \langle \boldsymbol{\gamma}(i), \mathbf{Z} \rangle$ and $J_1^*(\overline{\mathbf{Z}}) = \max_{i \in [k]} \langle \boldsymbol{\gamma}(i), \overline{\mathbf{Z}} \rangle = \langle \boldsymbol{\gamma}(\bar{i}_1^*), \overline{\mathbf{Z}} \rangle$, where $\bar{i}_1^* \in \arg\max_{i \in [k]} \langle \boldsymbol{\gamma}(i), \overline{\mathbf{Z}} \rangle$. Thus, we have:

$$J_1^*(\mathbf{Z}) = \langle \boldsymbol{\gamma}(i_1^*), \mathbf{Z} \rangle \geqslant \langle \boldsymbol{\gamma}(\bar{i}_1^*), \mathbf{Z} \rangle \overset{\text{(a)}}{\geqslant} \langle \boldsymbol{\gamma}(\bar{i}_1^*), \overline{\mathbf{Z}} \rangle = J_1^*(\overline{\mathbf{Z}}),$$

where inequality (a) follows from Assumption 1.a.

Suppose the statement holds for $T - 1$, we prove it for $T > 1$. To this end, we consider the *transition operator* $P : \mathcal{Z} \times [k] \times \mathbb{R} \to \mathcal{Z}$, defined for every observations vector $\mathbf{Z}_t = (1, X_{t-1}, X_{t-2}, \dots, X_{t-k}) \in \mathcal{Z}$, action $i \in [k]$, and noise $\xi \in \mathbb{R}$ as follows:

$$P(\mathbf{Z}_t, i, \xi) = P\left(\begin{pmatrix} 1 \\ X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-k} \end{pmatrix}, i, \xi\right) = \begin{pmatrix} 1 \\ X_t \\ X_{t-1} \\ \vdots \\ X_{t-k+1} \end{pmatrix} = \mathbf{Z}_{t+1}, \qquad \text{where} \qquad X_t = \langle \boldsymbol{\gamma}(i), \mathbf{Z}_t \rangle + \xi.$$

Thus, we can look at the stochastic process as a Markov decision process (Puterman, 2014) with

$\mathbf{Z}_t$ as state representation. We immediately observe that if $\mathbf{Z} \succeq \overline{\mathbf{Z}}$, we have that $P(\mathbf{Z}, i, \xi) \succeq P(\overline{\mathbf{Z}}, i, \xi)$, for every action $i \in [k]$ and noise $\xi \in \mathbb{R}$. By applying the Bellman equation, we obtain:

$$J_T^*(\mathbf{Z}) = \max_{i\in[k]} \left\{\langle \boldsymbol{\gamma}(i), \mathbf{Z}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{Z}, i, \xi_T))\right]\right\} = \langle \boldsymbol{\gamma}(i_T^*), \mathbf{Z}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{Z}, i_T^*, \xi_T))\right],$$
$$J_T^*(\overline{\mathbf{Z}}) = \max_{i\in[k]} \left\{\langle \boldsymbol{\gamma}(i), \overline{\mathbf{Z}}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{Z}}, i, \xi_T))\right]\right\} = \langle \boldsymbol{\gamma}(\bar{i}_T^*), \overline{\mathbf{Z}}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{Z}}, \bar{i}_T^*, \xi_T))\right],$$

where the actions are defined as $i_T^* \in \arg\max_{i\in[k]} \left\{\langle \boldsymbol{\gamma}(i), \mathbf{Z}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{Z}, i, \xi_T))\right]\right\}$ and $\bar{i}_T^* \in \arg\max_{i\in[k]} \left\{\langle \boldsymbol{\gamma}(i), \overline{\mathbf{Z}}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{Z}}, i, \xi_T))\right]\right\}$. Thus, we have:

$$\begin{aligned} J_T^*(\mathbf{Z}) &= \langle \boldsymbol{\gamma}(i_T^*), \mathbf{Z}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{Z}, i_T^*, \xi_T))\right] \\ &\geqslant \langle \boldsymbol{\gamma}(\bar{i}_T^*), \mathbf{Z}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\mathbf{Z}, \bar{i}_T^*, \xi_T))\right] \\ &\overset{(b)}{\geqslant} \langle \boldsymbol{\gamma}(\bar{i}_T^*), \overline{\mathbf{Z}}\rangle + \mathbb{E}_{\xi_T}\left[J_{T-1}^*(P(\overline{\mathbf{Z}}, \bar{i}_T^*, \xi_T))\right] = J_T^*(\overline{\mathbf{Z}}), \end{aligned}$$

where (b) follows from Assumption 1.a when bounding $\langle \boldsymbol{\gamma}(\bar{i}_T^*), \mathbf{Z}\rangle \geqslant \langle \boldsymbol{\gamma}(\bar{i}_T^*), \overline{\mathbf{Z}}\rangle$ and by observing that $P(\mathbf{Z}, \bar{i}_T^*, \xi_1) \succeq P(\overline{\mathbf{Z}}, \bar{i}_T^*, \xi_T)$ and, then, exploiting the inductive hypothesis.

We conclude that the optimal policy is the myopic one by observing that both $\langle \boldsymbol{\gamma}(i), \boldsymbol{z}\rangle$ and $J_{T-1}^*(P(\mathbf{Z}, i, \xi))$ are simultaneously maximized by $\arg\max_{i\in[k]}\langle \boldsymbol{\gamma}(i), \boldsymbol{z}\rangle$. ■

Lemma 2 (Self-Normalized Concentration)**.** *Let $i \in [k]$ be an action, let $\{\widehat{\boldsymbol{\gamma}}_t(i)\}_{t\in\mathcal{O}_\infty(i)}$ be the sequence of solutions to the Ridge regression problems computed by Algorithm 4. Then, for every regularization parameter $\lambda > 0$, confidence $\delta \in (0, 1)$, simultaneously for every round $t \in [T]$ and action $i \in [k]$, with probability at least $1 - \delta$ it holds that:*

$$\|\widehat{\boldsymbol{\gamma}}_t(i) - \boldsymbol{\gamma}(i)\|_{\mathbf{V}_t(i)} \leqslant \sqrt{\lambda}\|\boldsymbol{\gamma}(i)\|_2 + \sigma\sqrt{2\log\left(\frac{n}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(i)}{\lambda^{k+1}}\right)}.$$

Proof We consider an action at a time; then, the final result is obtained with a union bound over $[k]$. Let $i \in [k]$. We first observe that the estimates of action a change only when a is pulled. Let $l \in \mathbb{N}$ be an index and let $t_l(i) \in \mathbb{N}$ be the round in which action a is pulled for the l-th time, *i.e.*,

$\{t_l(i) \,:\, l \in \mathbb{N}\} = \mathcal{O}_\infty(i)$. Thus, we have:

$$
\begin{aligned}
\boldsymbol{\gamma}_{t_l}(i) &= \mathbf{V}^{-1}_{t_l(i)}(i)\mathbf{b}^{-1}_{t_l(i)}(i) \\
&= \left(\lambda \mathbf{I}_{m+1} + \sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1}\mathbf{Z}^T_{t_j(i)-1}\right)^{-1} \sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1} X_{t_j} \\
&= \left(\lambda \mathbf{I}_{m+1} + \sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1}\mathbf{Z}^T_{t_j(i)-1}\right)^{-1} \sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1} \left(\langle \boldsymbol{\gamma}(i), \mathbf{Z}_{t_j(i)-1}\rangle + \xi_{t_j(i)}\right) \\
&\overset{\text{(a)}}{=} \boldsymbol{\gamma}(i) - \lambda \left(\lambda \mathbf{I}_{m+1} + \sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1}\mathbf{Z}^T_{t_j(i)-1}\right)^{-1} \boldsymbol{\gamma}(i) + \\
&\qquad + \left(\lambda \mathbf{I}_{m+1} + \sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1}\mathbf{Z}^T_{t_j(i)-1}\right)^{-1} \sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1}\xi_{t_j(i)} \\
&= \boldsymbol{\gamma}(i) - \lambda \mathbf{V}^{-1}_{t_l(i)}(i)\boldsymbol{\gamma}(i) + \mathbf{V}^{-1}_{t_l(i)}(i) \underbrace{\sum_{j=1}^{l} \mathbf{Z}_{t_j(i)-1}\xi_{t_j(i)}}_{\mathbf{s}_{t_l(i)}},
\end{aligned}
$$

where the passage (a) derives from the observation that

$$
\sum_{j=1}^{l} \mathbf{Z}_{t_j-1}(\langle \boldsymbol{\gamma}(i), \mathbf{Z}_{t_j-1}\rangle) = \sum_{j=1}^{l} \mathbf{Z}_{t_j-1}\mathbf{Z}^T_{t_j-1}\boldsymbol{\gamma}(i)
$$

. Thus, we have:

$$
\left\|\boldsymbol{\gamma}_{t_l(i)}(i) - \boldsymbol{\gamma}(i)\right\|_{\mathbf{V}_{t_l(i)}(i)} \leqslant \sqrt{\lambda}\|\boldsymbol{\gamma}(i)\|_2 + \|\mathbf{s}_{t_l(i)}\|_{\mathbf{V}^{-1}_{t_l(i)}(i)}.
$$

Let us denote with $\mathcal{F}_{t_l(i)} = \sigma(\mathbf{Z}_0, i_1, \mathbf{Z}_1, i_2, \ldots, \mathbf{Z}_{t_l(i)-1}, i_{t_l(i)})$ be the filtration generated by all events realized at round $t_l(i)$. Let us now consider the stochastic processes $(\xi_{t_l(i)})_{l\in\mathbb{N}}$ and $(\mathbf{Z}_{t_l(i)-1})_{l\in\mathbb{N}}$. We observe that $\xi_{t_l(i)}$ is $\mathcal{F}_{t_l(i)}$-measurable and conditionally σ^2-subgaussian and that $\mathbf{Z}_{t_l(i)-1}$ is $\mathcal{F}_{t_l(i)-1}$-measurable. By applying Theorem 1 of Abbasi-Yadkori et al. (2011), we have that simultaneously for all $l \in \mathbb{N}$, w.p. $1-\delta$:

$$
\|\mathbf{s}_{t_l(i)}\|_{\mathbf{V}^{-1}_{t_l(i)}(i)} \leqslant \sigma\sqrt{2\log\frac{1}{\delta} + \log\frac{\det \mathbf{V}_{t_l(i)}(i)}{\lambda^{m+1}}}.
$$

Clearly, this hold for the rounds $t \in \mathbb{N}$ in which the action a is not pulled, since the corresponding estimates do not change. ■

Lemma 3 (Policy Regret Decomposition)**.** *Let* $(x^*_t)_{t\in[T]}$ *be the sequence of rewards by executing*

the optimal policy $\boldsymbol{\pi}^*$ and let $(X_t)_{t\in[T]}$ be the sequence of rewards by executing the learner's policy $\boldsymbol{\pi}$. Then, for every $t \in [T]$ it holds that:

$$\begin{aligned} r_t &= X_t^* - X_t \\ &= \sum_{j=1}^{m} \gamma_j(i_t^*)(X_{t-i}^* - X_{t-i}) + \langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(I_t), \mathbf{Z}_{t-1} \rangle \\ &= \sum_{j=1}^{m} \gamma_j(i_t^*) r_{t-i} + \rho_t, \end{aligned} \tag{3.9}$$

where $r_t \coloneqq X_t^* - X_t$ is the instantaneous policy regret, $\rho_t \coloneqq \langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(a_t), \mathbf{Z}_{t-1} \rangle$ is the instantaneous external regret, $i_t^* = \pi_t^*(H_{t-1}^*)$, and $r_{t-l} = 0$ if $l \geqslant t$.

Proof Let $t \in [T]$ and let us denote with $\mathbf{Z}_{t-1}^* = (1, X_{t-1}^*, \dots, X_{t-m}^*)^T$ the observations vector associated with the execution of the optimal policy and with $\mathbf{Z}_{t-1} = (1, X_{t-1}, \dots, X_{t-m})^T$ the observations vector associated with the execution of the learner's policy. We have:

$$\begin{aligned} r_t &= X_t^* - X_t \\ &= \langle \boldsymbol{\gamma}(i_t^*), \mathbf{Z}_{t-1}^* \rangle - \langle \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1} \rangle \\ &= \langle \boldsymbol{\gamma}(i_t^*), \mathbf{Z}_{t-1}^* \rangle - \langle \boldsymbol{\gamma}(i_t^*), \mathbf{Z}_{t-1} \rangle + \langle \boldsymbol{\gamma}(i_t^*), \mathbf{Z}_{t-1} \rangle - \langle \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1} \rangle \\ &= \langle \boldsymbol{\gamma}(i_t^*), \mathbf{Z}_{t-1}^* - \mathbf{Z}_{t-1} \rangle + \langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1} \rangle \\ &= \sum_{j=1}^{m} \gamma_j(i_t^*) \underbrace{(X_{t-i}^* - X_{t-i})}_{r_{t-i}} + \underbrace{\langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1} \rangle}_{\rho_t}, \end{aligned}$$

where in expanding the inner product we made the summation start from $j = 1$ as the two vectors $\mathbf{Z}_{t-1}^*$ and $\mathbf{Z}_{t-1}$ have the same first component equal to 1. ■

Lemma 4 (External-to-Policy Regret Bound)**.** *Let $\boldsymbol{\pi}$ be the learner's policy and $T \in \mathbb{N}$ be the horizon. Under Assumptions 1.a and 1.b, it holds that:*

$$\begin{aligned} R(\boldsymbol{\pi}, T) &= \mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{j=1}^{m} \gamma_j(i_t^*) r_{t-j} + \rho_t\right]\right] \\ &\leqslant \left(\frac{\Gamma m}{1-\Gamma} + 1\right) \varrho(\boldsymbol{\pi}, T), \end{aligned} \tag{3.10}$$

where $\varrho(\boldsymbol{\pi}, T) \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \rho_t\right]$ is the cumulative expected external regret.

Proof We start from the decomposition of Lemma 3. To prove the result we employ the so-called

"superposition principle", which allows us to decompose the linear recurrence as follows:

$$r_t = \sum_{j=1}^{m} \boldsymbol{\gamma}_j(i_t^*) r_{t-i} + \rho_t = \sum_{\tau=0}^{+\infty} \rho_\tau \widetilde{r}_{t,\tau},$$

where if $\tau > t$ we set $\widetilde{r}_{t,\tau} = 0$ and if $\tau \leqslant t$ we have that $\widetilde{r}_{t,\tau}$ is given by the recurrence:

$$\widetilde{r}_{t,\tau} = \sum_{j=1}^{m} \boldsymbol{\gamma}_j(i_t^*) \widetilde{r}_{t-i,\tau} + \delta_{t,\tau} \qquad \text{where} \qquad \delta_{t,\tau} := \begin{cases} 1 & t = \tau \\ 0 & t \neq \tau \end{cases}.$$

This way, we decompose the exogenous term ρ_τ as a linear combination of unitary impulses. Then by Assumption 1.a and 1.b, recalling that $\widetilde{r}_{t,\tau} = 0$ if $\tau > t$ and that $\widetilde{r}_{\tau,\tau} = 1$, we have that for every $t > \tau$ it holds that:

$$\widetilde{r}_{t,\tau} \leqslant \Gamma \max_{j \in [m]} \widetilde{r}_{t-j,\tau} \leqslant \Gamma^2 \max_{j \in [m]} \max_{l \in [m]} \widetilde{r}_{t-j-l,\tau} \leqslant \cdots \leqslant \Gamma^{\lceil (t-\tau)/m \rceil},$$

since we will encounter the $1 = \delta_{\tau,\tau}$ after $\lceil (t-\tau)/m \rceil$ steps of unfolding.

Now, we can manipulate this formula to have an expression of the full regret:

$$\begin{aligned}
\sum_{t=1}^{T} r_t &\leqslant \sum_{t=1}^{T} \left(\rho_t + \sum_{\tau=1}^{t-1} \Gamma^{\lceil (t-\tau)/m \rceil} \rho_\tau \right) \\
&= \sum_{\tau=1}^{T} \left(1 + \rho_\tau \sum_{t=\tau+1}^{T} \Gamma^{\lceil (t-\tau)/m \rceil} \right) \\
&\overset{(a)}{\leqslant} \sum_{\tau=1}^{T} \rho_\tau \left(1 + \sum_{s=1}^{+\infty} \Gamma^{\lceil s/m \rceil} \right) \\
&\overset{(b)}{=} \sum_{\tau=1}^{T} \rho_\tau \left(1 + \sum_{l=1}^{+\infty} m \Gamma^l \right) \\
&= \left(1 + \frac{\Gamma m}{1-\Gamma} \right) \sum_{\tau=1}^{T} \rho_\tau,
\end{aligned}$$

where (a) follows from bounding the summation with the series and changing the index $s = t - \tau$ and (b) is obtained by observing that the exponent $\lceil s/m \rceil$ changes only when s is divisible by m. ■

Counterexample to show that this bound is tight.

There are k arms:

$$\boldsymbol{\gamma}(i_1) := [\Gamma^0 \dots 0], \qquad \boldsymbol{\gamma}(i_2) := [0, \Gamma, 0 \dots 0], \;\; \dots \; \boldsymbol{\gamma}(i_k) := [0, \dots 0, \Gamma].$$

All these arms have non-negative coefficients whose sum is bounded by Γ. If the sequence of internal regrets is:

$$\rho_t = \begin{cases} 1 & t = 1 \\ 0 & t > 1 \end{cases},$$

and the sequence of arms is $i_1^* = 1$, and $i_t^* = i_{t-1 \ (mod\ k)+1}$ (which means $i_1, i_2, \dots, i_k, i_1, i_2, \dots$), we have:

$$r_1 = 1, r_2 = \Gamma, \; r_3 = \Gamma, \; \dots, \; r_{k+1} = \Gamma,$$

and then, we start again with the same sequence of arms:

$$r_{k+2} = \Gamma^2, \; r_{k+3} = \Gamma^2, \; \dots, \; r_{2k+1} = \Gamma^2.$$

Making the sum of these terms for t from one to infinity, we get:

$$\sum_{t=1}^{\infty} r_t = 1 + k \sum_{t=1}^{\infty} \Gamma^t = 1 + \frac{k\Gamma}{1-\Gamma},$$

which is exactly the bound we get.

Lemma 33. *Let $(\mathbf{Z}_t)_{t\in[T]}$ be the sequence of observation vectors observed by executing the learner's policy. If $\mathbf{Z}_0 = (1, 0, \dots, 0)^T$, then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$, simultaneously for all $t \in [T]$, it holds that:*

$$\|\mathbf{Z}_{t-1}\|_2 \leqslant \sqrt{1 + m \left(\frac{g+\eta}{1-\Gamma}\right)^2},$$

where $\eta = \sqrt{2\sigma^2 \log(T/\delta)}$.

Proof Let $(\xi_t)_{t\in[T]}$ be the sequence of noises. We consider the event $\mathcal{E} = \bigcap_{t=1}^{T} \{|\xi_t| \leqslant \eta\}$ prescribing that all noises are smaller than η in absolute value. By union bound, knowing that all the noises are independent σ^2-subgaussian random variables we, can bound the probability of

event $\mathcal{E}$:

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}\left(\bigcap_{t=1}^{T}\left\{|\xi_t| \leqslant \eta\right\}\right) \geqslant 1 - Te^{-\frac{\eta^2}{2\sigma^2}} = 1 - \delta,$$

having set $\eta = \sqrt{2\sigma^2 \log(T/\delta)}$. Under event $\mathcal{E}$ and when $\mathbf{Z}_0 = (1, 0, \dots, 0)^T$, we prove by induction that all rewards X_t are bounded in absolute value by $\frac{g+\eta}{1-\Gamma}$, regardless the actions played. For $T = 1$, the statement is trivial since $x_1 = \gamma_0(i_1) + \eta_1$ and, thus, $|x_1| \leqslant \gamma_0(i_1) + |\eta_1| \leqslant g + \eta \leqslant \frac{g+\eta}{1-\Gamma}$. Suppose the statement holds for all $s < t$, we prove it for t. We have:

$$\begin{aligned} X_t = \gamma_0(i_t) + \sum_{j=1}^{k} \gamma_j(i_t) X_{t-i} + \eta_t \quad \implies |X_t| &\leqslant \gamma_0(i_t) + \sum_{j=1}^{k} \gamma_j(i_t)|X_{t-i}| + |\eta_t| \\ &\leqslant g + \Gamma\frac{g+\Gamma}{1-\Gamma} + \eta = \frac{g+\eta}{1-\Gamma}, \end{aligned}$$

where the first inequality uses Assumption 1.a, the second inequality follows from the inductive hypothesis and by Assumptions 1.b and 1.c. Passing to the observations vector, we have:

$$\|\mathbf{Z}_{t-1}\|_2^2 = 1 + \sum_{j=1}^{m} X_{t-i}^2 \leqslant 1 + m\left(\frac{g+\eta}{1-\Gamma}\right)^2.$$

■

For deriving the regret bound, we make use of the following result, known as *Elliptic Potential Lemma* (Lattimore and Szepesvári, 2020, Lemma 19.4).

Lemma 34 (Elliptic Potential Lemma)**.** *Let $\mathbf{V}_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix and let $\boldsymbol{i}_1, \dots, \boldsymbol{i}_n \in \mathbb{R}^d$ be a sequence of vectors such that $\|\boldsymbol{i}_t\|_2 \leqslant L < +\infty$ for all $t \in [k]$. Let $\mathbf{V}_t = \mathbf{V}_0 + \sum_{s=1}^{t} \boldsymbol{i}_s \boldsymbol{i}_s^T$, Then:*

$$\sum_{t=1}^{n} \min\{1, \|\boldsymbol{i}_s\|_{\mathbf{V}_{t-1}^{-1}}\} \leqslant 2d \log\left(\frac{\mathrm{tr}(\mathbf{V}_0) + nL^2}{d\det(\mathbf{V}_0)^{1/d}}\right).$$

Theorem 5. *Let $\delta = (2T)^{-1}$. Under Assumptions 1.a, 1.b, and 1.c,* AR−UCB *suffers a cumulative expected (policy) regret bounded by (highlighting the dependence on g, σ, m, Γ, k, and T):*

$$\mathbb{E}[R(\texttt{AR-UCB}, T)] \leqslant \widetilde{\mathcal{O}}\left(\frac{(g+\sigma)(m+1)^{3/2}\sqrt{kT}}{(1-\Gamma)^2}\right).$$

Proof We denote with $(X_t^*)_{t\in[T]}$ the sequence of rewards generated by playing the optimal

policy and with $(X_t)_{t\in[T]}$ the sequence of rewards generated by playing AR−UCB. Thanks to Lemma 4, we have to bound the external regret only. Let $\delta \in (0,1)$, and define, as in the main paper, for every round $t \in [T]$ and action $i \in [k]$:

$$\beta_t(i) := \sqrt{\lambda(g^2+1)} + \sigma\sqrt{2\log\left(\frac{k}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_t(i)}{\lambda^{m+1}}\right)}.$$

Let us define the confidence set $\mathcal{C}_t(i) := \{\boldsymbol{\gamma} \in \mathbb{R}^{m+1} : \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{t-1}(i)\|_{\mathbf{V}_{t-1}(i)} \leqslant \beta_{t-1}(i)\}$ and the optimistic estimate of the true parameter vector $\boldsymbol{\gamma}(i)$:

$$\widetilde{\boldsymbol{\gamma}}_t(i) \in \operatorname*{arg\,max}_{\boldsymbol{\gamma}\in\mathcal{C}_t(i)} \langle \boldsymbol{\gamma}, \mathbf{Z}_{t-1}\rangle,$$

By Theorem 2, we have that, for every action $i \in [k]$ and round $t \in [T]$, the true parameter vector satisfies $\boldsymbol{\gamma}(i) \in \mathcal{C}_t(i)$ with probability at least $1-\delta$. Therefore, with the same probability, we have:

$$\begin{aligned}
\langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1}\rangle &= \underbrace{\langle \boldsymbol{\gamma}(i_t^*) - \widetilde{\boldsymbol{\gamma}}_t(i_t), \mathbf{Z}_{t-1}\rangle}_{\leqslant 0} + \langle \widetilde{\boldsymbol{\gamma}}_t(i_t) - \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1}\rangle \\
&\leqslant \langle \widetilde{\boldsymbol{\gamma}}_t(i_t) - \widehat{\boldsymbol{\gamma}}_{t-1}(i_t), \mathbf{Z}_{t-1}\rangle + \langle \widehat{\boldsymbol{\gamma}}_{t-1}(i_t) - \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1}\rangle \\
&\leqslant 2\beta_{t-1}(i_t)\|\mathbf{Z}_{t-1}\|_{\mathbf{V}_{t-1}(i)^{-1}},
\end{aligned}$$

where the first inequality follows from the optimism and in the last passage we have used Cauchy-Schwartz inequality, recalling that for every couple of vectors $\boldsymbol{v}, \boldsymbol{w}$ it holds $\langle \boldsymbol{v}, \boldsymbol{w}\rangle \leqslant \|\boldsymbol{v}\|_{\mathbf{V}_{t-1}(i)}\|\boldsymbol{w}\|_{\mathbf{V}_{t-1}(i)^{-1}}$, and having observed that $\boldsymbol{\gamma}(i_t), \widetilde{\boldsymbol{\gamma}}_t(i_t) \in \mathcal{C}_t(i_t)$.

Furthermore, we observe that the external regret $\rho_t = \langle \boldsymbol{\gamma}(i_t^*) - \boldsymbol{\gamma}(i_t), \mathbf{Z}_{t-1}\rangle \leqslant \|\boldsymbol{z}_{t-1}\|_2 + g$, since the coefficients γ_j for $j \neq 0$ have a sum bounded by $\Gamma < 1$ and get multiplied by $\mathbf{Z}_{t-1}$, while γ_0, which is bounded by g gets multiplied by 1, then we have $\rho_t \leqslant L + g = \mathcal{O}(g)$. By Lemma 33 with probability of at least $1-\delta$ we have:

$$\|\boldsymbol{z}_t\|_2 \leqslant \sqrt{1 + m\left(\frac{g+\eta}{1-\Gamma}\right)^2} =: L,$$

where $\eta = \sqrt{2\sigma^2\log(T/\delta)}$ and, consequently:

$$\rho_t \leqslant g + L =: C_1.$$

At this point, we proceed as follows:

$$\rho_t \leqslant 2\min\{C_1, \beta_{t-1}(i_t)\|\mathbf{Z}_{t-1}\|_{\mathbf{V}_{t-1}(i_t)^{-1}}\} \leqslant 2\max\{C_1, \beta_{t-1}(i_t)\}\min\{1, \|\mathbf{Z}_{t-1}\|_{\mathbf{V}_{t-1}(i_t)^{-1}}\}.$$

Summing over $t \in [T]$, we obtain a bound on the cumulative external regret:

$$\begin{aligned}\varrho(\texttt{AR-UCB}, T) = \sum_{t=1}^{T} \rho_t &= \sum_{t=1}^{T} 1 \cdot \rho_t \\ &\leqslant \sqrt{T \sum_{t=1}^{T} \rho_t^2} \\ &\leqslant 2 \max\{C_1, \beta_{T-1}\} \sqrt{T \sum_{t=1}^{T} \min\{1, \|\mathbf{Z}_{t-1}\|^2_{\mathbf{V}_{t-1}(i_t)^{-1}}\}}\end{aligned}$$

where:

$$\beta_{T-1} := \max_{a \in [k]} \beta_{T-1}(i),$$

where the first inequality follows from an application of Cauchy-Schwartz inequality and the last passage holds since the sequence $\beta_t(i_t)$ is non-decreasing, and so we can bound each of them with their value at $t = T$. Now, we are finally able to use the *Elliptic Potential Lemma* (Lemma 34):

$$\begin{aligned}\sum_{t=1}^{T} \min\{1, \|\mathbf{Z}_{t-1}\|^2_{\mathbf{V}_{t-1}(i_t)^{-1}}\} &= \sum_{i \in [k]} \sum_{l \in \mathcal{O}_T(i)} \min\{1, \|\mathbf{Z}_{l-1}\|^2_{\mathbf{V}_{l-1}(i)^{-1}}\} \\ &\leqslant \sum_{i \in [k]} 2(m+1) \log\left(\frac{\lambda(m+1) + |\mathcal{O}_T(i)| L^2}{\lambda(m+1)}\right) \\ &\leqslant 2k(m+1) \log\left(1 + \frac{TL^2}{k\lambda(m+1)}\right),\end{aligned}$$

where the first inequality follows from an application of the elliptic potential lemma for each action $i \in [k]$ observing that $\mathbf{V}_0 = \lambda \mathbf{I}_{m+1}$ and, consequently, $\text{tr}(\mathbf{V}_0) = \lambda(m+1)$ and $\det(\mathbf{V}_0)^{1/(m+1)} = \lambda$. The second inequality follows by observing that $\sum_{i \in [k]} |\mathcal{O}_T(i)| = T$ and since the $\log$ is a concave function, the worst allocation of pulls is the uniform one. Now that we have bounded the inner summation, we can state that:

$$\varrho(\texttt{AR-UCB}, T) = \sum_{t=1}^{T} \rho_t \leqslant 2 \max\{C_1, \beta_{T-1}\} \sqrt{2Tk(m+1) \log\left(1 + \frac{TL^2}{k\lambda(m+1)}\right)}.$$

To conclude, we bound the term β_{T-1} as follows:

$$\begin{aligned}\beta_{T-1} &= \sqrt{\lambda(g^2+1)} + \sigma \max_{a\in[k]} \sqrt{2\log\left(\frac{k}{\delta}\right) + \log\left(\frac{\det \mathbf{V}_{T-1}(i)}{\lambda^{m+1}}\right)} \\ &\leqslant \sqrt{\lambda(g^2+1)} + \sigma\sqrt{2\log\left(\frac{k}{\delta}\right) + (m+1)\log\left(\frac{\lambda(m+1)+TL^2}{\lambda(m+1)}\right)}.\end{aligned}$$

Therefore, by highlighting the dependences on g, m, σ, and Γ, we have:

$$\beta_{T-1} = \widetilde{O}\left(g + \sigma\sqrt{m+1}\right), \qquad C_1 = \widetilde{\mathcal{O}}\left(1 + \sqrt{m}\frac{m+\sigma}{1-\Gamma}\right).$$

These results hold with probability $1-2\delta$. We set $\delta = (2T)^{-1}$. Putting all together, we obtain:

$$\varrho(\texttt{AR-UCB}, T) = \sum_{t=1}^{T} \rho_t \leqslant \widetilde{\mathcal{O}}\left(\frac{(g+\sigma)\sqrt{k(m+1)T}}{1-\Gamma}\right),$$

and, applying the previous Lemma 4, this results in:

$$R(\texttt{AR-UCB}, T) \leqslant \widetilde{\mathcal{O}}\left(\frac{(g+\sigma)(m+1)^{3/2}\sqrt{kT}}{(1-\Gamma)^2}\right).$$

■

A.2. Optimal Policy without Noise

In the case of no noise, our system writes:

$$X_t = \gamma_0(i_t) + \sum_{j=1}^{m} \gamma_j(i_t) X_{t-i}. \tag{A.1}$$

In this case, the process evolution is deterministic. Therefore, even if it is still true that the optimal policy is given by Theorem 1, it is possible to say that there is a constant policy that is asymptotically optimal, in the sense that its cumulative regret is bounded by a constant. This policy is given by:

$$i^* \in \arg\max_{a\in\mathcal{A}} \frac{\gamma_0(i_t)}{1-\sum_{j=1}^{m}\gamma_j(i_t)}. \tag{A.2}$$

This result is not surprising. In fact, this action makes the process converge to the highest

possible stationary reward, which is of course $\arg\max_{a\in\mathcal{A}} \frac{\gamma_0(i_t)}{1-\sum_{j=1}^m \gamma_j(i_t)}$. Precisely, the following result holds.

Theorem 35. *Let us consider the problem formulation of Equation* (A.1)*. Define:*

$$i^* = \arg\max_{a\in\mathcal{A}} \frac{\gamma_0(i_t)}{1-\sum_{j=1}^m \gamma_j(i_t)},$$

as in Equation (A.2)*. Then, there exist no policy $\boldsymbol{\pi}$ (even non-constant) such that:*

$$\limsup_{t\to+\infty} X_t^{\boldsymbol{\pi}} - X_t^* > 0$$

(where $X_t^{\boldsymbol{\pi}}$ denotes the sequence obtained with policy $\boldsymbol{\pi}$, while X_t^* is the one relative to i^*). Moreover, the cumulative regret with respect to the actual optimal policy is bounded by:

$$\gamma_0(i^*)\frac{m}{(1-\Gamma)^2}.$$

Proof If we play always i^*, we have:

$$\limsup_{t\to+\infty} X_t^* = \frac{\gamma_0(i^*)}{1-\sum_{j=1}^m \gamma_j(i^*)},$$

by imposing the condition of stationarity. For the rest of the proof, let us denote:

$$X^* := \frac{\gamma_0(i^*)}{1-\sum_{j=1}^m \gamma_j(i^*)}.$$

Now, we prove that, for any policy $\boldsymbol{\pi}$, we cannot achieve an $X_t > X^*$. By contradiction, if $\limsup_{t\to\infty} X_t^{\boldsymbol{\pi}} - X_t^* > 0$, then the set $\{t \in \mathbb{N} \,:\, X_t > X^*\}$ is non-empty. Let $t_0 = \min\{t \in \mathbb{N} \,:\, X_t > X^*\}$. Then, by definition:

$$X_{t_0} = \gamma_0(i_{t_0}) + \sum_{j=1}^m \gamma_j(i_{t_0}) x_{t_0-i}.$$

Recalling that t_0 is the first time in which we surpass X^*, we have:

$$X^* < X_{t_0} = \gamma_0(i_{t_0}) + \sum_{j=1}^m \gamma_j(i_{t_0}) x_{t_0-i} \leqslant \gamma_0(i_{t_0}) + \sum_{j=1}^m \gamma_j(i_{t_0}) X^*.$$

This inequality entails that:

$$\Big(1 - \sum_{j=1}^{m} \gamma_j(i_{t_0})\Big) X^* < \gamma_0(i_{t_0}),$$

and, therefore:

$$\frac{\gamma_0(i^*)}{1 - \sum_{j=1}^{m} \gamma_j(i^*)} = X^* < \frac{\gamma_0(i_{t_0})}{1 - \sum_{j=1}^{m} \gamma_j(i_{t_0})},$$

which contradicts the definition of i^*.

For the second part, we start considering that the regret obtained by using constant action i^* is bounded by:

$$\sum_{t=1}^{+\infty} X^* - X_t,$$

since X^* is the maximum instantaneous reward that every policy can achieve. Now, note that $\gamma_0(i^*) > 0$, otherwise it could not be the optimal action. At this point, we have for $0 < t \leqslant m$ that $X_t \geqslant \gamma_0(i^*)$, by simply using the fact that all the coefficients of the autoregressive model are non-negative. From this fact we have for $m < t \leqslant 2m$ that $X_t \geqslant \gamma_0(i^*)(1 + \sum_{j=1}^{m} \gamma_j(i^*))$; and generalizing:

$$\forall j > 0 \qquad \text{and} \qquad jm - m < t \leqslant jm: \qquad X_t \geqslant \gamma_0(i^*)\Big(\sum_{\ell=0}^{j} (\Gamma^*)^\ell\Big), \qquad \Gamma^* = \sum_{j=1}^{m} \gamma_j(i^*).$$

Therefore, we have $X_t \geqslant \gamma_0(i^*)\frac{1-\Gamma^{\lfloor t/m \rfloor}}{1-\Gamma}$, which means:

$$\begin{aligned}
R_t &\leqslant \sum_{t=1}^{+\infty} X^* - X_t \\
&\leqslant \sum_{t=1}^{+\infty} X^* - \gamma_0(i^*)\frac{1-\Gamma^{\lfloor t/m \rfloor}}{1-\Gamma} \\
&= \gamma_0(i^*) \sum_{t=1}^{+\infty} \frac{1}{1-\Gamma} - \frac{1-\Gamma^{\lfloor t/m \rfloor}}{1-\Gamma} \\
&= \gamma_0(i^*) \sum_{t=1}^{+\infty} \frac{\Gamma^{\lfloor t/m \rfloor}}{1-\Gamma} \\
&= \gamma_0(i^*) \frac{m}{(1-\Gamma)^2}.
\end{aligned}$$

■

A.3. Discussion on Assumption 1.a

In this appendix, we further detail the meaning of Assumption 1.a related to non-negative coefficients governing the AR process (Assumption 1.a). Even if, theoretically, this setting is less general than the one that considers all possible values for the parameters, we argue that, for the real-world applications of interest, considering negative coefficients is not meaningful.

Before introducing our example, let us remark on the meaning of X_t in practice. This value represents the sales volume in the case of pricing, the value of a stock in the stock market, the number of customers that an e-commerce website may have, and so on. In all these real-world scenarios, the quantity X_t is meaningful whenever we consider non-negative values that we want to maximize. We argue that when Assumption 1.a is not fulfilled (*i.e.*, at least one $\gamma_j(i)$ is negative), the positivity of X_t is no longer ensured.

Consider now the example presented in Figure A.1, where we present a general scenario in which, at time τ, we are in a given with a certain positive x_τ. Consider, for the sake of simplicity, a noiseless setting with $m = 1$ (*i.e.*, an AR(1) process) and, for a given action i, we have $\gamma_0(i) = 0$. Consider now $\gamma_1(i) < 0$. Figure A.1 shows what will happen in this case. The value of X_t continuously changes its sign at each time step, and this behavior is not compatible with the real-world phenomena of our interest. This is even more unrealistic if we think about the scenario in which we have another value of the state $\overline{x}_\tau > x_\tau$. In this scenario, after performing the same action i, we will observe that the best-starting state $\overline{x}_\tau$ leads to a worst next state $\overline{x}_{\tau+1} < x_{\tau+1}$. This behavior has no practical meaning in the applications of our interest. Given these considerations, we can derive that the worst possible effect of a given action is to *reset* the state, which corresponds to have $\gamma_1(i) = 0$. A representation of this phenomenon is drawn in Figure A.2. From this figure, it is possible to notice how a process can always decrease as an effect of an action, even for $\gamma_1(i) > 0$.

This consideration trivially generalizes for any $m > 1$ given a generic state representation $\mathbf{Z}_\tau$.

A.4. Additional Experimental Results

In this appendix, we provide additional experimental results. In Appendix A.4.1, we assert the effectiveness of `AR-UCB` in the classic stochastic bandit problem by comparing its performances with two standard baselines from the literature. In Appendix A.4.2, we stress the effect of mis-specifying parameter $\overline{m}$ in the standard multi-armed bandit problem. Finally, in Appendix A.4.3, we provide experimental results in the particular case of autoregressive processes of order 1 (*i.e.*, $m = 1$).

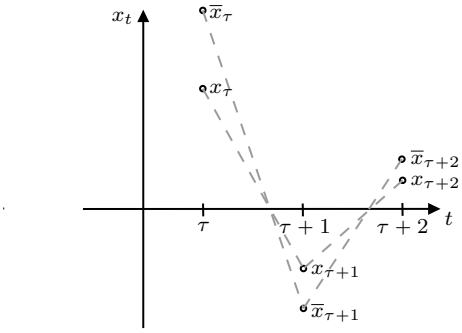


Figure A.1: An illustration of the effect of a negative $\gamma_1(i)$ over time.

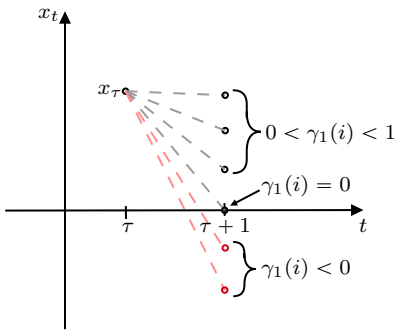
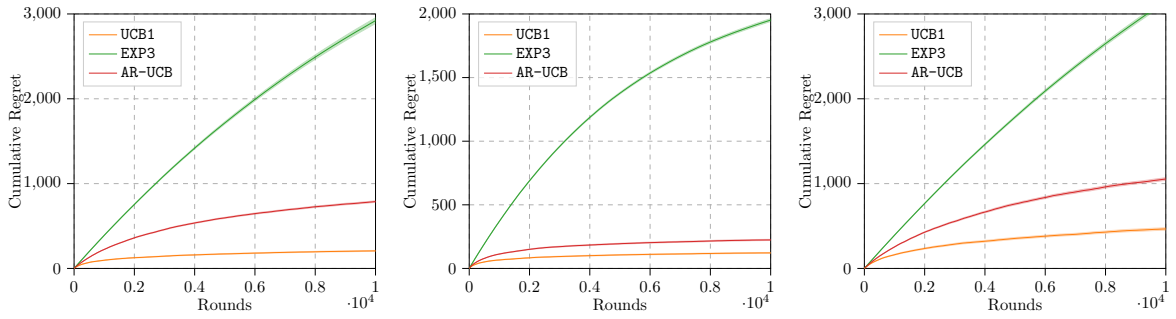


Figure A.2: The effect of $\gamma_1(i)$ in the evolution of the state X_t, in the case of a non-negative one (in black), and a negative one (in red).



(a) Setting A. (b) Setting B. (c) Setting C.

Figure A.3: Cumulative regret of AR−UCB, UCB1, and EXP3 in the case of $k = 0$ (100 runs, mean $\pm$ std).

A.4.1. Stochastic Bandit Problem

Setting We evaluate AR−UCB in the special case $m = 0$. This problem is equivalent to solving a standard stochastic bandit problem. This experiment compares the performances of AR−UCB in this setting against well-known gold standards: UCB1 and EXP3. The competing algorithms are evaluated in terms of cumulative regret w.r.t. the setting-specific clairvoyant. The three settings differ in the values of $g \in \{2, 7.5\}$ (*i.e.*, the maximum arms' expected reward) and the values of $\sigma \in \{0.9, 1.25, 2\}$, the noise's standard deviation.

Results Figure A.3 shows the average cumulative regrets for AR−UCB, UCB1, and EXP3. We immediately observe that all the algorithms suffer sublinear cumulative regret, as expected since they are all able to provide no-regret theoretical guarantees in this setting. In all the experiments, UCB1 outperforms all the other algorithms since it is specifically designed for the scenario under analysis. AR−UCB, as expected, performs properly in this setting since, as already discussed in

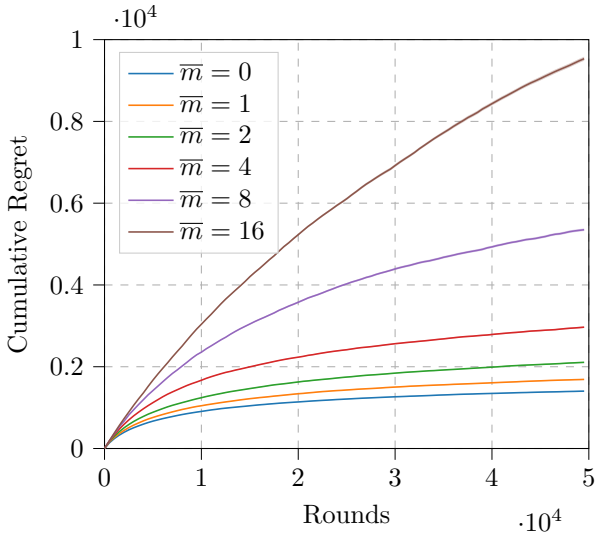


Figure A.4: Cumulative regret of AR−UCB in the case of $m = 0$, in with $\overline{m}$ parameter misspecified (100 runs, mean $\pm$ std).

Section 3.1.5, its regret is asymptotically optimal when $m = 0$.

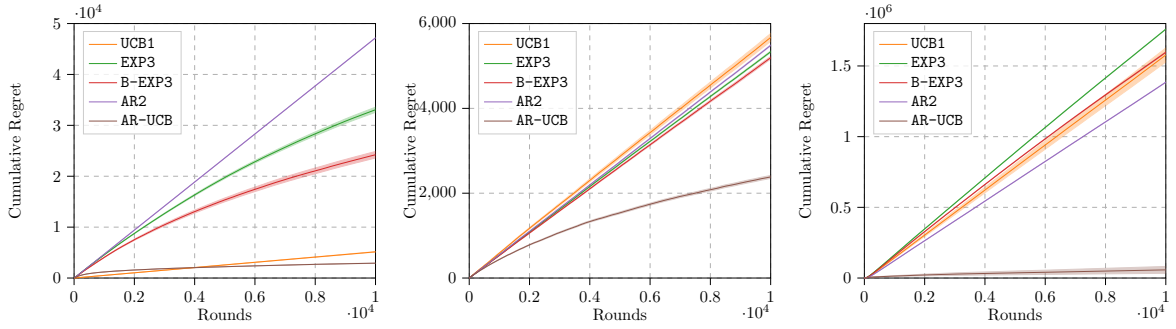
A.4.2. On the Misspecification of k in Stochastic Bandit Problem

Setting We evaluate AR−UCB in the special case $m = 0$. This problem is equivalent to solving a standard stochastic bandit problem. This experiment compares the performances of AR−UCB under different values of the parameter $\overline{k}$. In particular, this experiment aims to highlight the performances of AR−UCB under a misspecification of the process memory in the special case where the true underlying process does not present a dynamic temporal structure. The parameters $\gamma_0(i)$ have been sampled by a uniform distribution having support $[6, 7]$, and g is set to 10. The noise's standard deviation σ is set to 1. The number of actions is $k = 7$.

Results Figure A.4 shows the average cumulative regrets for AR−UCB under different values of $\overline{m}$, when the true value is $m = 0$. The figure shows that AR−UCB is capable of achieving sublinear cumulative regret even when the misspecification is severe (*e.g.*, $\overline{m} = 16$), consistent with the theoretical results; the performance degrades as the misspecification grows.

A.4.3. AR(1) Bandit Problem

AR(1) processes are the simplest autoregressive processes. Therefore, we will present a specific analysis of this setting to show how AR−UCB and the baselines perform when the complexity given by the dynamic temporal structure is minimal. Results show how even the minimal autoregressive contribution can lead all the algorithms (except for AR−UCB) to linear cumulative



(a) Setting A. (b) Setting B. (c) Setting C.

Figure A.5: Cumulative regret of AR−UCB and the others bandit baselines in the case of $m = 1$ (100 runs, mean $\pm$ std).

regret.

Setting We evaluate AR−UCB in the case $m = 1$. This is the simplest setting in which an autoregressive component contributes to the reward. This experiment compares the performances of AR−UCB in this setting against the same baselines as Section 3.1.6. The competing algorithms are evaluated in terms of cumulative regret w.r.t. the setting-specific clairvoyant. The three settings differ in the values of $g \in \{2, 8, 10\}$ (*i.e.*, the maximum arms' expected reward) and the values of $\sigma \in \{1, 1.25, 2\}$, the noise's standard deviation. The values of the $\gamma_1(i)$ parameters have been sampled from uniform distributions having their sampling ranges inside $[0, 1)$. The number of actions is $k = 7$.

Results Figure A.5 shows the average cumulative regrets for all the competing algorithms. We immediately observe that the only algorithms able to achieve sublinear regret are AR−UCB (in all three settings), B−EXP3 (first and third experiments), and EXP3 (first experiment only). Such a result is unsurprising since none of the baselines has specific theoretical guarantees in the Autoregressive Bandit problem, even in the simple scenario when $m = 1$. Even though, we decided to adopt these algorithms as baselines since they represent the gold standard algorithms in the bandit literature (UCB1, EXP3) and the algorithms that solve problems near to ours (B−EXP3 , AR2), respectively.

B | Graph-Triggered Bandits

B.1. Proofs on Rising Bandits

In this appendix, we report a short version of the proofs of Rising GTBs. The extended version is provided in (Genalti et al., 2024c).

Theorem 6 (Complexity of finding the Optimal Policy in Rising GTBs)**.** *Compu- ting the optimal policy in Rising GTBs with general matrices* $\mathbf{G}$ *is NP-Hard.*

Proof We reduce from a decision problem related to finding cliques in graphs. In particular, given a graph (V, E) and $\widetilde{M} \in \mathbb{N}$, it is NP-Hard to determine if there exists a clique of size $\widetilde{M}$ (Karp, 1972). In the following, we design an instance of our problem such that the reward of the optimal policy is at least $\sum_{t=1}^{T}(1 + \frac{t}{T^2})$ if and only if there exists a clique of size $\widetilde{M} = T$.

Construction. Given a graph (V, E), we build an instance such that the horizon is T. Our set of actions can be constructed by assigning an action to every node and time step couple, *i.e.*, $\mathcal{A} = \{a_{v,t}\}_{v \in V,\ t \in [T]}$. We define the matrix $\widetilde{\mathbf{G}}$ is such that for any $v, v' \in V$ and $t, t' \in [T]$, it holds $G_{a_{v,t},a_{v',t'}} = 1$ if $(v, v') \in E$, and $G_{a_{v,t},a_{v',t'}} = 0$ otherwise. Finally, for each arm $a_{v,t} \in \mathcal{A}$, the reward is deterministic and evolves as $\mu_{a_{v,t}}(n) = \min\{1 + \eta t, \frac{n}{t}(1 + \eta t)\}$, where $\eta = T^{-2}$. We call $\widetilde{\boldsymbol{\nu}}$ the set of these functions. It is easy to see that the GTB instance $(\widetilde{\boldsymbol{\nu}}, \widetilde{\mathbf{G}}, T)$ satisfies assumption 2.

if. We show that if there exists a clique $C^\star = \{v_1, \dots, v_T\}$ of size T, then there exists a policy with a cumulative reward of at least $\sum_{t=1}^{T}(1 + \eta t)$. Consider the policy $\widetilde{\pi}$ s.t. $\widetilde{\pi}(t) = a_{v_t,t}$. It is easy to see that $\widetilde{N}_{a_{v_t,t},t} = t$ for every $t \in [T]$. Hence, the reward of the policy $\widetilde{\pi}$ at time t is

$$\mu_{a_{v_t,t}}(\widetilde{N}_{a_{v_t,t},t}) = \min\{1 + \eta t, \frac{t}{t}(1 + \eta t)\} = 1 + \eta t.$$

Thus, $J_{\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{G}},T}(\widetilde{\pi}) = \sum_{t=1}^{T}(1 + \eta t)$ and the claim is proven.

only if. We show that if there is a policy $\widetilde{\pi}$ s.t. $J_{\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{G}},T}(\widetilde{\pi}) \geqslant \sum_{t=1}^{T}(1 + \eta t)$, then there exists a clique of size T.

First, we observe that for each $t', t \in [T]$ it holds that

$$\max_{t' \in [T]} \min\{1 + \eta t', \frac{t}{t'}(1 + \eta t')\} = 1 + \eta t. \tag{B.1}$$

This implies that, at any round t, the best obtainable reward is

$$\begin{aligned}\max_{t' \in [T]} \max_{v \in V} \max_{l \leqslant t} \mu_{a_{v,t'}}(l) &= \max_{t' \in [T]} \max_{v \in V} \widetilde{\mu}_{a_{v,t'}}(t) \\ &= \max_{t' \in [T]} \min \left\{ 1 + \eta t', \frac{t}{t'}(1 + \eta t') \right\} \\ &= \min \left\{ 1 + \eta t, \frac{t}{t}(1 + \eta t) \right\} = 1 + \eta t.\end{aligned}$$

Since by assumption there is a policy with reward at least $\sum_{t=1}^{T}(1 + \eta t)$, then there is a policy such that at each round $t \in [T]$ the reward is exactly $1 + \eta t$.

Consider a round $t \in [T]$. Let $a_{v,t'}$ be the arm played by the policy at this round. It must be the case that: i) $t' = t$, otherwise

$$\mu_{a_{v,t'}}(\widetilde{N}_{a_{v,t'},t}) \leqslant \mu_{a_{v,t'}}(t) < 1 + \eta t$$

by Equation (B.1), and ii) $\widetilde{N}_{a_{v,t'},t} = t$, otherwise

$$\mu_{a_{v,t'}}(\widetilde{N}_{a_{v,t'},t}) \leqslant \frac{t-1}{t}(1 + \eta t) < 1 + \eta t.$$

Let $a_{v_t,t}$ be the arm chosen at round t. Then, each arm in $\{a_{v_t,t}\}_{t \in [T]}$, is chosen while having exactly $t-1$ triggers. By the definition of $\widetilde{\mathbf{G}}$ this directly implies that $\{v_t\}_{t=1}^{T}$ is a clique of size T. ■

Theorem 7 (Optimal Policy in Rising GTBs with Block-Diagonal CM)**.** *For any instance $(\boldsymbol{\nu}, \mathbf{G}, T)$ of Rising GTBs with $\mathbf{G} \in \mathbb{B}_{\tilde{k}}$, the optimal policy $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j \in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_j(t), \qquad \forall t \in [T],$$

where $C^*_{\boldsymbol{\nu},\mathbf{G},T}$ is the "best" cumulative reward clique:

$$C^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_{C \in \mathcal{C}_{\mathbf{G}}} \sum_{t \in [T]} \max_{j \in C} \mu_j(t).$$

Proof For each clique $C_m \in \mathcal{C}_{\mathbf{G}}$, we substitute the reward function of every arm $i \in C_m$ with $\mu_i^*(t) = \max_{i \in C_m} \mu_i(t)$, for every $t \in [T]$. Now, since all arms sharing the same clique have the same reward function, our instance is equivalent to a $\widetilde{k}$-armed bandit problem, where $\widetilde{k}$ is the number of cliques. Since arms in different cliques are not connected, this corresponds to a rested bandit problem, and we use Proposition 1 of (Heidari et al., 2016) to get that the optimal policy would only pull the best action in terms of cumulative reward at the end of the time horizon T. To conclude the proof, we remark that playing greedily inside a clique corresponds exactly to play on the reward function defined above, which dominates the initial problem, and so the maximum cumulative reward is exactly the one attained in the problem with $\widetilde{k}$ arms. ■

Lemma 36 (DR−BD−UB Estimator's Instantaneous Bias)**.** *For every arm $i \in [k]$, every round $t > 1$, let us define:*

$$\bar{\mu}_i(t) \coloneqq \mu_i(t^I_{i,N_{i,t-1}}) + (t - t^I_{i,N_{i,t-1}}) \frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}},$$

then, $\bar{\mu}_i(t) \geqslant \mu_i(t^I_{i,N_{i,t-1}})$ and, if $N_{i,t-1} \geqslant 2$ it holds that:

$$\bar{\mu}_i(t) - \mu_i(\widetilde{N}_{i,t}) \leqslant (t - t^I_{i,N_{i,t}-1})\gamma_i(t^I_{i,N_{i,t-1}-1}).$$

Proof Let us start by observing the following equality holding:

$$\mu_i(\widetilde{N}_{i,t}) = \mu_i(t^I_{i,N_{i,t-1}}) + \sum_{j=t^I_{i,N_{i,t-1}}}^{\widetilde{N}_{i,t}-1} \gamma_i(j).$$

We have:

$$\begin{aligned}
\mu_i(\widetilde{N}_{i,t}) &= \mu_i(t^I_{i,N_{i,t-1}}) + \sum_{j=t^I_{i,N_{i,t-1}}}^{\widetilde{N}_{i,t}-1} \gamma_i(j) \\
&\leqslant \mu_i(t^I_{i,N_{i,t-1}}) + (\widetilde{N}_{i,t} - t^I_{i,N_{i,t-1}})\gamma_i(t^I_{i,N_{i,t-1}-1}) && \text{(B.2)} \\
&\leqslant \mu_i(t^I_{i,N_{i,t-1}}) + (t - t^I_{i,N_{i,t-1}})\gamma_i(t^I_{i,N_{i,t-1}-1}), && \text{(B.3)}
\end{aligned}$$

where line (B.2) follows from Assumption 2, and line (B.3) is obtained from observing that

$\widetilde{N}_{i,t} \leqslant t$. Concerning the bias, when $N_{i,t-1} \geqslant 2$, we have:

$$\bar{\mu}_i(t) - \mu_i(\widetilde{N}_{i,t}) \leqslant \mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(\widetilde{N}_{i,t}) + (t - t^I_{i,N_{i,t-1}}) \frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} \tag{B.4}$$

$$\leqslant (t - t^I_{i,N_{i,t-1}}) \frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} \tag{B.5}$$

$$\leqslant (t - t^I_{i,N_{i,t-1}}) \gamma_i(t^I_{i,N_{i,t-1}-1}), \tag{B.6}$$

where line (B.5) follows from observing that $\mu_i(t^I_{i,N_{i,t-1}}) \leqslant \mu_i(\widetilde{N}_{i,t})$, and line (B.6) derives from bounding $\frac{\mu_i(t^I_{i,N_{i,t-1}}) - \mu_i(t^I_{i,N_{i,t-1}-1})}{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-1}} \leqslant \gamma_i(t^I_{i,N_{i,t-1}-1})$ thanks to Assumption 2. ■

Theorem 8 (DR-BD-UB Regret in Det. Rising GTBs with Block-Diagonal CMs)**.** *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$ and $\sigma = 0$. Then,* DR-BD-UB *suffers a regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) \leqslant \widetilde{\mathcal{O}}\Bigg(\inf_{q \in [0,1]} \Bigg\{ \underbrace{T^q \sum_{C_m \in \mathcal{C}} |C_m| \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)}_{\text{(A) Rested Bias Contribution}} + \underbrace{\sum_{C_m \in \mathcal{C}} |C_m| \widetilde{N}^{\frac{q}{1+q}}_{C_m,T} \Upsilon_{\boldsymbol{\nu}} \left(\left\lceil \frac{\widetilde{N}_{C_m,T}}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}}}_{\text{(B) Restless Bias Contribution}} \Bigg\} \Bigg).$$

Proof Let $C^*_{\boldsymbol{\nu},\mathbf{G},T} \in \mathcal{C}_{\mathbf{G}}$ be the optimal clique of the instance. We analyze the following expression:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) = \sum_{t=1}^{T} \mu_{i^*_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}),$$

where $i^*_t \in \arg\max_{i \in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_i(t)$ for all $t \in [T]$. Then, we can decompose the regret in two

meaningful components:

$$
\begin{aligned}
R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-BD-UB}) &= \sum_{t=1}^{T} \mu_{i_t^*}(t) \pm \bar{\mu}_{I_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t}) \\
&\leqslant \sum_{t=1}^{T} \min\{1, \bar{\mu}_{I_t}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\} && \text{(B.7)} \\
&\leqslant \sum_{t=1}^{T} \min\{1, (t - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} && \text{(B.8)} \\
&= \sum_{t=1}^{T} \min\{1, (t \pm t^I_{I_t,N_{I_t,t}} - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} \\
&\leqslant \sum_{t=1}^{T} \min\{1, (t - t^I_{I_t,N_{I_t,t}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} + && \text{(B.9)} \\
&\quad + \sum_{t=1} \min\{1, (t^I_{I_t,N_{I_t,t}} - t^I_{I_t,N_{I_t,t-1}})\gamma_{I_t}(t^I_{I_t,N_{I_t,t-1}-1})\} && \text{(B.10)} \\
&= 4k + \underbrace{\sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} \min\{1, (t - t^I_{i,j})\gamma(t^I_{i,j-2})\}}_{\text{(a)}} + \\
&\quad + \underbrace{\sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{i \in C_m} \sum_{j=3}^{N_{j,T}} \min\{1, (t^I_{i,j} - t^I_{i,j-1})\gamma(t^I_{i,j-2})\}}_{\text{(b)}},
\end{aligned}
$$

where lines (B.7) and (B.8) follow from Lemma 36, line (B.10) from the fact that $\min\{1, x+y\} \leqslant \min\{1, x\} + \min\{1, y\}$ for any $x, y \geqslant 0$.

These two terms represent the rested and the restless contribution to the regret, and we can bound them using similar techniques as in (Metelli et al., 2022). ■

Remark 13 (Regret Bound in Rested and Restless Rising Bandits)**.** *When we are in a purely rested (resp. restless) scenario, the contribution term associated to the restless (resp. rested) scenario vanish, and we get the same regret orders from (Metelli et al., 2022). In particular, we can avoid splitting the minimum in Equation* (B.10) *and instead notice that in a rested setting we have* $t - t^I_{I_t,N_{I_t,t-1}} = t - N_{I_t,t-1}$*, and thus we can bound the cumulative regret as we bound the term (a). Instead, in a restless setting we have* $t - t^I_{I_t,N_{I_t,t-1}} = t - t_{I_t,N_{I_t,t-1}}$*, and thus we can bound the cumulative regret as we bound the term (b).*

Theorem 9 (DR-G-UB Regret in Det. Rising GTBs with General Matrices)**.** *Let*

$(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \{0,1\}^{k\times k}$ and $\sigma = 0$. Then, DR-G-UB *suffers a regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{DR-G-UB})$$
$$\leqslant \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{T^q \sum_{C_m^L\in\mathcal{C}_{\bar{\mathbf{G}}^L}} |C_m^L|\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil\frac{\widetilde{N}_{C_m^L,T}}{|C_m^L|}\right\rceil, q\right) + \sum_{C_m^L\in\mathcal{C}_{\bar{\mathbf{G}}^L}} |C_m^L|\widetilde{N}_{C_m^L,T}^{\frac{q}{1+q}}\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil\frac{\widetilde{N}_{C_m^L,T}}{|C_m^L|}\right\rceil, q\right)^{\frac{1}{1+q}}\right\}\right),$$

where $\bar{\mathbf{G}}^L \in \mathbb{B}_{\widetilde{k}}$ is a maximal sub-matrix of $\mathbf{G}$.

Proof The theorem can be proved by showing that estimator's bias is always larger when internal times are decreased. For every arm $i \in [k]$ we define:

$$f_i(t;\ x, y) = \mu_i(x) + (t-x)\frac{\mu_i(x) - \mu_i(y)}{x-y}, \tag{B.11}$$

for every triplet of natural numbers $y \leqslant x \leqslant t \leqslant T$. Note that $\bar{\mu}_i(t) = f_i(t;\ t^I_{i,N_{i,t-1}}, t^I_{i,N_{i,t-1}-1})$, so if we can show that f_i is decreasing in both x and y, we can prove the claim. We start with the second argument: fix t and x, then for any y:

$$\begin{aligned} f_i(t;x,y) - f_i(t;x,y-1) &= (t-x)\left(\frac{\sum_{j=y}^{x-1}\gamma_i(j)}{x-y} - \frac{\sum_{j=y-1}^{x-1}\gamma_i(j)}{x-y+1}\right) \\ &= \frac{\sum_{j=y}^{x-1}\gamma_i(j) - (x-y)\gamma_i(y-1)}{(x-y)(x-y+1)} \leqslant 0, \end{aligned} \tag{B.12}$$

where line (B.12) follows from Assumption 2. With slightly more calculations we show that f_i is also decreasing in the first argument, fix t and y, then for any x:

$$f_i(t;\ x,y) - f_i(t;\ x-1,y) \leqslant 0. \tag{B.13}$$

Now we observe that, for every $i \in [k]$ and every $t \in [T]$, we have $t^I_{i,N_{i,t}} \geqslant t^{I,L}_{i,N_{i,t}}$. This is a consequence of Definition 1, since:

$$t^I_{i,N_{i,t}} - t^{I,L}_{i,N_{i,t}} = \sum_{j=1}^{t}(G_{I_t,i} - \bar{G}^L_{I_t,i}) \geqslant 0.$$

As a consequence of this, we have:

$$f_i(t;\ t^I_{i,N_{i,t-1}}, t^I_{i,N_{i,t}-1}) \leqslant f_i(t;\ t^{I,L}_{i,N_{i,t-1}}, t^{I,L}_{i,N_{i,t}-1}), \tag{B.14}$$

and

$$\mu_i(t^I_{i,N_{i,t}}) \geqslant \mu_i(t^{I,L}_{i,N_{i,t}}). \tag{B.15}$$

The proof can be concluded in the same way as for Theorem 8. ■

Lemma 37 (Estimator's Instantaneous Bias)**.** *For every arm $i \in [k]$, every round $t \in [T]$, and window width $1 \leqslant h \leqslant \left\lfloor \frac{N_{i,t-1}}{2} \right\rfloor$, let us define:*

$$\widetilde{\mu}^h_i(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + (t-l)\frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h} \right),$$

otherwise if $h = 0$, we set $\widetilde{\mu}^h_i(t) \coloneqq +\infty$. Then, $\widetilde{\mu}^h_i(t) \geqslant \mu_i(t_{i,N_{i,t-1}})$ and, if $N_{i,t-1} \geqslant 2$ it holds that:

$$\widetilde{\mu}^h_i(t) - \mu_i(\widetilde{N}_{i,t}) \leqslant \frac{(2t - 2N_{i,t-1} + h - 1)(t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1})}{2h} \gamma_i(t^I_{i,N_{i,t-1}-2h+1}).$$

Proof Let us start by observing the following equality holding for every $l \in \{2, \dots, N_{i,t-1}\}$:

$$\mu_i(\widetilde{N}_{i,t}) = \mu_i(t^I_{i,l}) + \sum_{j=t^I_{i,l}}^{\widetilde{N}_{i,t}-1} \gamma_i(j).$$

By averaging over a window of length h, we obtain:

$$\begin{aligned}
\mu_i(\widetilde{N}_{i,t}) &= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + \sum_{j=t^I_{i,l}}^{\widetilde{N}_{i,t}-1} \gamma_i(j) \right) \\
&\leqslant \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + (\widetilde{N}_{i,t} - t^I_{i,l})\gamma_i(t^I_{i,l} - 1) \right) && \text{(B.16)} \\
&\leqslant \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + \frac{\widetilde{N}_{i,t} - t^I_{i,l}}{t^I_{i,l} - t^I_{i,l-h}} \sum_{j=t^I_{i,l-h}}^{t^I_{i,l}-1} \gamma_i(j) \right) && \text{(B.17)} \\
&\leqslant \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + (t-l)\frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h} \right) =: \widetilde{\mu}^h_i(t), && \text{(B.18)}
\end{aligned}$$

where lines (B.16) and (B.17) follow from Assumption 2, and line (B.18) is obtained from observing that $t^I_{i,l} \geqslant l$, $\widetilde{N}_{i,t} \leqslant t$ and $t^I_{i,l} - t^I_{i,l-h} \geqslant h$.

Concerning the bias, when $N_{i,t-1} \geqslant 2$, we have:

$$\begin{aligned}
\widetilde{\mu}_i^h(t) - \mu_i(\widetilde{N}_{i,t}) &= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t^I_{i,l}) + (t-l)\frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h} \right) - \mu_i(\widetilde{N}_{i,t}) \\
&\leqslant \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l)\frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{h} && \text{(B.19)} \\
&= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l)\frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{t^I_{i,l} - t^I_{i,l-h}} \frac{t^I_{i,l} - t^I_{i,l-h}}{h} \\
&\leqslant \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l)\gamma_i(t^I_{i,l-h})\frac{t^I_{i,l} - t^I_{i,l-h}}{h} && \text{(B.20)} \\
&\leqslant \frac{t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1}}{h^2}\gamma_i(t^I_{i,N_{i,t-1}-2h+1}) \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l) && \text{(B.21)} \\
&= \frac{(2t - 2N_{i,t-1} + h - 1)(t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1})}{2h}\gamma_i(t^I_{i,N_{i,t-1}-2h+1}), && \text{(B.22)}
\end{aligned}$$

where line (B.19) follows from observing that $\mu_i(t^I_{i,l}) \leqslant \mu_i(\widetilde{N}_{i,t})$, line (B.20) derives from Assumption 2 and bounding $\frac{\mu_i(t^I_{i,l}) - \mu_i(t^I_{i,l-h})}{t^I_{i,l} - t^I_{i,l-h}} \leqslant \gamma_i(t^I_{i,l-h})$, line (B.21) is obtained by bounding $t^I_{i,l} - t^I_{i,l-h} \leqslant t^I_{i,N_{i,t-1}} - t^I_{i,N_{i,t-1}-2h+1}$ and $\gamma_i(t^I_{i,l-h}) \leqslant \gamma_i(t^I_{i,N_{i,t-1}-2h+1})$, and line (B.22) follows from computing the summation.

■

Lemma 38 (Bound on Estimator's Cumulative Bias for Block-Diagonal CMs). *Let $(I_t)_{t=1}$ be a sequence of actions. For every action $i \in [k]$, every round $t \in [T]$, let window width $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$. Let $\mathbf{G} \in \mathbb{B}^{\tilde{k}}$ be a block diagonal matrix, then for every $q \in [0,1]$, we have:*

$$\begin{aligned}
&\sum_{t=1}^{T} \min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\} \leqslant \\
&\quad \leqslant 2k + \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{\bar{k}_1} \right\rceil, q \right) + \\
&\quad + T^{\frac{2q}{1+q}} (1 + \log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil \sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m| > 1} |C_m| \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{|C_m|} \right\rceil, q \right)^{\frac{1}{1+q}},
\end{aligned}$$

where $\mathcal{C}$ is the set of blocks of matrix $\mathbf{G}$, and $\bar{k}_1 \leqslant k$ is the number of blocks of size 1.

Proof The statement can be proven by decomposing over the cliques and then over the arms,

splitting cliques with only one arm from the others:

$$\sum_{t=1}^{T} \min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\} \leqslant 2k + \underbrace{\sum_{\substack{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m|=1 \\ C_m=\{i\}}} \sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(j)\right\}}_{\text{(a)}} + \underbrace{\sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m|>1} \sum_{i \in C_m} \sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(t_{i,j}^I)\right\}}_{\text{(b)}}.$$

The two terms can be bound in a similar way as in (Metelli et al., 2022), as the rested and the restless component, respectively. ■

Theorem 11 (R-▫-UCB Regret in Rising GTBs with Block-Diagonal CMs)**.** *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \mathbb{B}_{\widetilde{k}}$. Let $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$ and $\delta_t = t^{-\alpha}$ for $\alpha > 2$. Then,* R-▫-UCB *suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\text{R-▫-UCB}) \leqslant \widetilde{\mathcal{O}}\left(\min_{q \in [0,1]} \left\{ \underbrace{(\sigma T)^{\frac{2}{3}}}_{\text{(A) Variance Contribution}} + \underbrace{\bar{k}_1 T^q \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{\bar{k}_1} \right\rceil, q\right)}_{\text{(B) Rested Bias Contribution}} + \underbrace{T^{\frac{2q}{1+q}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}} : |C_m|>1} |C_m| \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{|C_m|} \right\rceil, q\right)^{\frac{1}{1+q}}}_{\text{(C) Restless Bias Contribution}} \right\} \right),$$

where $\bar{k}_1$ is the number of cliques in $\mathbf{G}$ containing only one action.

Proof Let us define the good events $\mathcal{E}_t = \bigcap_{i \in [k]} \mathcal{E}_{i,t}$ that correspond to the event in which all confidence intervals hold:

$$\mathcal{E}_{i,t} := \left\{ \left| \widehat{\mu}_i^{h_{i,t}}(t) - \widetilde{\mu}_i^{h_{i,t}}(t) \right| \leqslant \beta_i^{h_{i,t}}(t) \right\} \qquad \forall i \in [T],\ i \in [k].$$

We have to analyze the following expression:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\text{DR-BD-UB}) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i_t^*}(t) - \mu_{I_t}(t) \right],$$

where $i_t^* \in \arg\max_{i \in C^*_{\boldsymbol{\nu},\mathbf{G},T}} \mu_i(t)$ for all $t = 1$. We decompose according to the good events $\mathcal{E}_t$:

$$\begin{aligned} R_{\boldsymbol{\nu},\mathbf{G},T}(\pi^{\texttt{DR-BD-UB}}) &= \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right) \mathbb{1}\{\mathcal{E}_t\}\right] + \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right) \mathbb{1}\{\neg\mathcal{E}_t\}\right] \\ &\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right) \mathbb{1}\{\mathcal{E}_t\}\right] + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\}\right], \end{aligned}$$

where we exploited $\mu_{i_t^*}(t) - \mu_{I_t}(t) \leqslant 1$ in the inequality. The second summation can be bounded using standard arguments, recalling that $\alpha > 2$:

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\}\right] &\leqslant 1 + \sum_{i \in [k]} \sum_{t=2}^{T} \mathbb{P}\left(\neg\mathcal{E}_{i,t}\right) \\ &\leqslant 1 + \frac{2k}{\alpha - 2}. \end{aligned}$$

where the first inequality is obtained with $\mathbb{P}(\neg\mathcal{E}_1) \leqslant 1$ and a union bound over $[k]$. Recalling $\mathbb{P}(\neg\mathcal{E}_{i,t})$ was bounded in Lemma 10, we bound the summation with the integral and obtain the second inequality.

The rest of the analysis can be conducted under the good event $\mathcal{E}_t$, recalling that $B_i(t) \equiv \widehat{\mu}_i^{h_{i,t}}(t) + \beta_i^{h_{i,t}}(t)$. Let $t \in [T]$, and we exploit the optimism, *i.e.*, $B_{i_t^*}(t) \leqslant B_{I_t}(t)$:

$$\begin{aligned} \mu_{i^*}(t) - \mu_{I_t}(t) + B_{I_t}(t) - B_{I_t}(t) &\leqslant \min\left\{1, \underbrace{\mu_{i_t^*}(t) - B_{i_t^*}(t)}_{\leqslant 0} + B_{I_t}(t) - \mu_{I_t}(t)\right\} \\ &\leqslant \min\left\{1, B_{I_t}(t) - \mu_{I_t}(t)\right\}. \end{aligned}$$

Now, we work on the term inside the minimum:

$$\begin{aligned} B_{I_t}(t) - \mu_{I_t}(t) &= \widehat{\mu}_{I_t}^{h_{I_t,t}}(t) + \beta_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t) && \text{(B.23)} \\ &\leqslant \underbrace{\widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(t)}_{\text{(a)}} + \underbrace{2\beta_{I_t}^{h_{I_t,t}}(t)}_{\text{(b)}}, && \text{(B.24)} \end{aligned}$$

where line (B.23) follows from the definition of $B_i(t)$ and line (B.24) from the good event $\mathcal{E}_t$. We make use of Lemma 38 and Lemma 42 to bound the summations over t of (a) and (b), respectively.

Putting all together, we obtain:

$$
\begin{aligned}
&R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-}\square\texttt{-UCB}) \\
&\quad\leqslant 1+\frac{2k}{\alpha-2}+5k+\frac{k}{\epsilon}+\frac{3k}{\epsilon}(2\sigma T)^{\frac{2}{3}}\left(10\alpha\log T\right)^{\frac{1}{3}}+ \\
&\qquad+T^{\frac{2q}{1+q}}(1+\log(\epsilon T))^{\frac{q}{1+q}}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil k\Upsilon_{\boldsymbol{\mu}}\left(\left\lceil(1-2\epsilon)\frac{T}{k}\right\rceil,q\right)^{\frac{1}{1+q}}+ \\
&\qquad+2k+\bar{k}_1T^q\left\lceil\frac{1}{1-2\epsilon}\right\rceil\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil(1-2\epsilon)\frac{T}{\bar{k}_1}\right\rceil,q\right)+ \\
&\qquad+T^{\frac{2q}{1+q}}(1+\log(\epsilon T))^{\frac{q}{1+q}}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{C_m\in\mathcal{C}_{\mathbf{G}}:|C_m|>1}|C_m|\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil(1-2\epsilon)\frac{T}{|C_m|}\right\rceil,q\right)^{\frac{1}{1+q}}.
\end{aligned}
$$

■

Lemma 39 (Bound on Estimator's Cumulative Bias for General Matrices)**.** *Let* $\{I_t\}_{t=1}$ *be a sequence of actions. For every action* $i\in[k]$*, every round* $t\in[T]$*, let window width* $h_{i,t}=\lfloor\epsilon N_{i,t-1}\rfloor$*. Let* $\mathbf{G}\in\{0,1\}^{k\times k}$*, then for every* $q\in[0,1]$*, we have*

$$
\begin{aligned}
&\sum_{t=1}^{T}\min\left\{1,\widetilde{\mu}_{I_t}^{h_{I_t},t}(t)-\mu_{I_t}(\widetilde{N}_{I_t,t})\right\}\leqslant \\
&\qquad\leqslant 2k+\bar{k}_1T^q\left\lceil\frac{1}{1-2\epsilon}\right\rceil\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil(1-2\epsilon)\frac{T}{\bar{k}_1}\right\rceil,q\right)+ \\
&\qquad\quad+T^{\frac{2q}{1+q}}(1+\log(\epsilon T))^{\frac{q}{1+q}}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil(k-\bar{k}_1)\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil(1-2\epsilon)\frac{T}{k-\bar{k}_1}\right\rceil,q\right)^{\frac{1}{1+q}},
\end{aligned}
\tag{B.25}
$$

where $\bar{k}_1\leqslant k$ *is the number of arms having degree of* 1, i.e., $\bar{k}_1:=|\{i\in[k]:deg(i)=1\}|$.

Proof The proof follows similar steps as Lemma 38. We decided to split arms based on their degree; in particular, we bound separately the bias due to arms having a degree of 1 (i.e., they are only triggered by themselves).

$$\sum_{t=1}^{T} \min\left\{1, \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t) - \mu_{I_t}(\widetilde{N}_{I_t,t})\right\}$$
$$\leqslant 2k + \underbrace{\sum_{\substack{i\in[k]\\ \deg^-(i)=1}} \sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(j)\right\}}_{\text{(a)}} + \underbrace{\sum_{\substack{i\in[k]\\ \deg^-(i)>1}} \sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_i^{h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(t^I_{i,j})\right\}}_{\text{(b)}}.$$

As a consequence of Definition 1, we observe that:

$$t^I_{i,N_{i,t}} - t^{I,U}_{i,N_{i,t}} = \sum_{j=1}^{t} (G_{I_t,i} - \bar{G}^U_{I_t,i}) \leqslant 0.$$

As a consequence of this, we have that, for every $i \in [k]$ and for every $t \in [T]$:

$$\widetilde{N}_{i,t} \leqslant \widetilde{N}^U_{i,t}, \tag{B.26}$$

where $\widetilde{N}^U_{i,t} := \mathbf{e}_i^\top (\bar{\mathbf{G}}^U)^\top \mathbf{N}_t$. Then, following similar steps as in (Metelli et al., 2022), we can bound the two components separately and make the dependency on the upper block-diagonal matrix explicit. ■

Theorem 12 (R−□−UCB Regret in Rising GTBs with General Matrices)**.** *Let $(\boldsymbol{\nu}, \mathbf{G}, T)$ be an instance of Rising GTB, where $\mathbf{G} \in \{0,1\}^{k\times k}$. Let $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$ and $\delta_t = t^{-\alpha}$ for $\alpha > 2$. Then,* R−□−UCB *suffers an expected regret bounded by:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\text{R−□−UCB}) \leqslant \widetilde{\mathcal{O}}\left(\min_{q\in[0,1]}\left\{(\sigma T)^{\frac{2}{3}} + T^q \bar{k}_1 \Upsilon_\nu\left(\frac{T}{\bar{k}_1}, q\right) + T^{\frac{2q}{1+q}} \sum_{C^U_m} |C^U_m| \Upsilon_\nu\left(\frac{T}{|C^U_m|}, q\right)^{\frac{1}{1+q}}\right\}\right),$$

where $\bar{\mathbf{G}}^U$ is the minimal super-matrix of $\mathbf{G}$.

Proof The proof follows similar steps of the proof of Theorem 11, but uses Lemma 39 (instead of Lemma 38) to bound cumulative estimator's bias.

As in Theorem 11, we decompose the regret in two components and instead make use of Lemma 39 and Lemma 42 to bound the summations over t of the two components, respectively.

Putting all together, we obtain:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{R-}\square\texttt{-UCB}) \leqslant 1 + \frac{2k}{\alpha-2} + 5k + \frac{k}{\epsilon} + \frac{3k}{\epsilon}(2\sigma T)^{\frac{2}{3}}\left(10\alpha\log T\right)^{\frac{1}{3}} + \\ + 2k + \bar{k}_1 T^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{\bar{k}_1}\right\rceil, q\right) + \\ + T^{\frac{2q}{1+q}}\left(1+\log(\epsilon T)\right)^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon}\right\rceil \left\lceil \frac{1}{1-2\epsilon}\right\rceil \cdot \sum_{\substack{C_m^U \in \mathcal{C}_{\bar{\mathbf{G}}^U} \\ |C_m^U|>1}} |C_m| \Upsilon_{\boldsymbol{\nu}}\left(\left\lceil (1-2\epsilon)\frac{T}{|C_m|}\right\rceil, q\right)^{\frac{1}{1+q}}.$$

■

B.1.1. Technical Lemmas

Lemma 40 (Lemma C.1 of Metelli et al. 2022)**.** *Let $M \geqslant 3$, and let $f : \mathbb{N} \to \mathbb{R}$, and $\beta \in (0,1)$. Then it holds that:*

$$\sum_{j=3}^{M} f(\lfloor \beta j \rfloor) \leqslant \left\lceil \frac{1}{\beta} \right\rceil \sum_{l=\lfloor 3\beta \rfloor}^{\lfloor \beta M \rfloor} f(l).$$

Lemma 41 (Lemma C.2 of Metelli et al. 2022)**.** *Under Assumption 2, it holds that:*

$$\max_{\substack{(N_{i,T})_{i\in[k]} \\ N_{i,T}\geqslant 0, \sum_{i\in[k]} N_{i,T} = T}} \sum_{i\in[k]} \sum_{l=1}^{N_{i,T}-1} \gamma_i(l)^q \leqslant k\Upsilon_{\boldsymbol{\nu}}\left(\left\lceil \frac{T}{k}\right\rceil, q\right).$$

Lemma 10 (Concentration of Estimator, adapted from Metelli et al. 2022)**.** *For every arm $i \in [k]$, every round $t \in [T]$, and window width $1 \leqslant h \leqslant \left\lfloor \frac{N_{i,t-1}}{2} \right\rfloor$, let:*

$$\beta_i^h(t,\delta) := \sigma(t - N_{i,t-1} + h - 1)\sqrt{\frac{10\log\frac{1}{\delta}}{h^3}}.$$

Then, if the window size depends on the number of pulls only $h_{i,t} = h(N_{i,t-1})$ and if $\delta_t = t^{-\alpha}$ for some $\alpha > 2$, it holds for every round $t \in [T]$ that:

$$\mathbb{P}\left(\left|\widehat{\mu}_i^{h_{i,t}}(t) - \widetilde{\mu}_i^{h_{i,t}}(t)\right| > \beta_i^{h_{i,t}}(t,\delta_t)\right) \leqslant 2t^{1-\alpha}.$$

Proof Using a Doob's *optional skipping* argument (Doob, 1953; Bubeck et al., 2008), and noting that, at round t, $t_{i,l}^I$ is a stopping time for every arm $i \in [k]$ and pull number $l \in \{1, \dots, N_{i,t-1}\}$

w.r.t. the filtration $\mathcal{F}_{\tau-1} = \sigma(I_1, X_1, \dots, I_{\tau-1}, X_{\tau-1}, I_\tau)$, we can proceed to prove this lemma as in (Metelli et al., 2022) also for GTB. ■

Lemma 42 (Bound on Estimator's Variance, Theorem 4.4 of Metelli et al. 2022)**.** *Let* $(I_t)_{t\in[T]}$ *be a sequence of actions such that:*

$$\left|\widehat{\mu}_{I_t}^{h_{I_t,t}}(t) - \widetilde{\mu}_{I_t}^{h_{I_t,t}}(t)\right| \leqslant \beta_{I_t}^{h_{I_t,t}}(t, t^{-\alpha}),\ \forall t \in [T], \tag{B.27}$$

where $\alpha > 2$*. For every action* $i \in [k]$*, every round* $t \in [T]$*, let window width* $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$*, then, we have:*

$$\sum_{t=1}^{T} \min\left\{1, 2\beta_{I_t}^{h_{I_t,t}}(t, t^{-\alpha})\right\} \leqslant k\left(3 + \frac{1}{\epsilon}\right) + \frac{3k}{\epsilon}(2\sigma T)^{\frac{2}{3}}(10\alpha \log T)^{\frac{1}{3}}. \tag{B.28}$$

B.2. Proofs on Rotting Bandits

Theorem 13 (Complexity of finding the Optimal Policy in Rotting GTBs)**.** *Computing the optimal policy in Rotting GTBs with general matrices* $\mathbf{G}$ *is NP-Hard.*

Proof We reduce from a decision problem related to finding independent sets in graphs. In particular, given a graph (V, E) and $\widetilde{M} \in \mathbb{N}$, it is NP-Hard to determine if there exists an independent set of size $\widetilde{M}$ (Karp, 1972). In the following, we design an instance of our problem such that the reward of the optimal policy is at least T if and only if there exists an independent set of size $\widetilde{M} = T$.

Construction. Given a graph (V, E), we build an instance such that the horizon is T. Our set of actions can be constructed by assigning an action to every node, *i.e.*, $\mathcal{A} = \{a_v\}_{v\in V}$. We define the matrix $\widetilde{\mathbf{G}}$ is such that for any $v, v' \in V$, it holds $G_{a_v,a_{v'}} = 1$ if $(v, v') \in E$, and $G_{a_v,a_{v'}} = 0$ otherwise. Finally, for each arm $a_v \in \mathcal{A}$, the reward is deterministic and evolves as $\mu_{a_{v,t}}(n) = \max\{2 - n, 0\}$. We call $\widetilde{\boldsymbol{\nu}}$ the set of these functions. It is easy to see that the GTB instance $(\widetilde{\boldsymbol{\nu}}, \widetilde{\mathbf{G}}, T)$ satisfies assumption 3.

if. We show that if there exists an independent set $I^\star = \{v_1, \dots, v_T\}$ of size T, then there exists a policy with a cumulative reward of at least T. Consider the policy $\widetilde{\pi}$ s.t. $\widetilde{\pi}(t) = a_{v_t}$. It is easy to see that $\widetilde{N}_{a_{v_t},t} = 1$ for every $t \in [T]$. Hence, the reward of the policy $\widetilde{\pi}$ at time t is

$$\mu_{a_{v_t}}(\widetilde{N}_{a_{v_t},t}) = 1.$$

Thus, $J_{\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{G}},T}(\widetilde{\pi}) = T$ and the claim is proven.

only if. We show that if there is a policy $\widetilde{\pi}$ s.t. $J_{\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{G}},T}(\widetilde{\pi}) \geqslant T$, then there exists an independent set of size T. First, we observe that at any round t the best obtainable reward is 1. Since, by assumption, there is a policy with a reward of at least T; then there is a policy such that at each round $t \in [T]$, the reward is exactly 1.

Let a_{v_t} be the arm played by the policy at round $t \in [T]$. Then, consider a round $t \in [T]$. Since the reward of the arm a_{v_t} must be 1, it must be the case that $\mu_{a_{v_t}}(\widetilde{N}_{a_{v_t},t}) = 1$ and $\widetilde{N}_{a_{v_t},t} = 1$. By the definition of $\widetilde{\mathbf{G}}$ this directly implies that $\{v_t\}$ is not connected to any $v_{t'}$, $t' < t$, and that $v_{t'} \neq v_t$ for any $t' < t$. Hence, $\{v_t\}_{t\in[T]}$ is an independent set of size T, proving the claim. ■

Lemma 43. *Let $\mathcal{G}_n = (V_n, E_n)$ be a graph. Then, either the graph possesses block-diagonal connectivity and the nodes can be partitioned in M disjoint clique, i.e., $V = \bigcup_{m=1}^{M} C_m$, or there exist three nodes v_1, v_2 and v_3 s.t. e_{v_1,v_2}, $e_{v_2,v_3} \in E$ and $e_{v_1,v_3} \notin E$.*

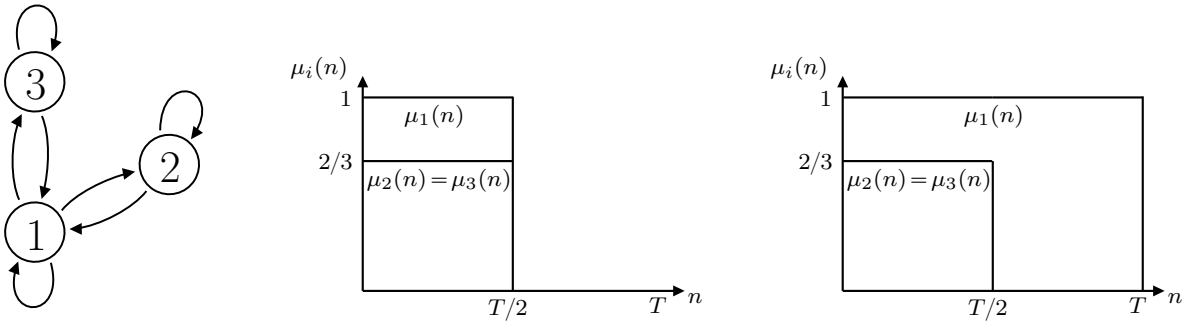
Proof We proceed by induction. The statement holds for $n \leqslant 3$. We assume that $\mathcal{G}_n$ is an arbitrary graph satisfying the statement. Then we add one node v_{n+1} and obtain $\mathcal{G}_{n+1} = (V_{n+1}, E_{n+1})$.

If $\mathcal{G}_n$ is block-diagonal connected. We list all the possible scenarios:

- If $e_{v_{n+1},i} \notin E_{n+1}$ for every $i \in V_n$ then the node v_{n+1} is a single-element clique and the new graph is block-diagonal.
- If $e_{v_{n+1},i} \in E_{n+1}$ for every $i \in C_m$, and $e_{v_{n+1},j} \notin E_{n+1}$ for every $j \in V_n \backslash C_m$, then the node v_{n+1} is added to the clique C_m and the new graph is block-diagonal.
- If $e_{v_{n+1},i} \in E_{n+1}$ for some $i \in C_m$ and $e_{v_{n+1},j} \notin E_{n+1}$ for some $j \in C_m$, then $e_{v_{n+1},i}$, $e_{i,j} \in E$ and $e_{v_{n+1},j} \notin E$.
- If $e_{v_{n+1},i} \in E_{n+1}$ for some $i \in C_m$ and $e_{v_{n+1},j} \in E_{n+1}$ for some $j \in C_{m'}$, then $e_{v_{n+1},i}$, $e_{v_{n+1},j} \in E$ and $e_{i,j} \notin E$.

If there exists three nodes v_1, v_2 and v_3 s.t. e_{v_1,v_2}, $e_{v_2,v_3} \in E$ and $e_{v_1,v_3} \notin E$. There is no way to connect v_1 and v_3 by adding a node, thus the statement still holds for $\mathcal{G}_{n+1}$. ■

Theorem 16 (Regret Lower Bound for Rotting GTBs with General Matrices)**.** *For every $\mathbf{G} \in \{0,1\}^{k\times k}$ that is not block-diagonal, there exists an instance of Rotting GTB $(\boldsymbol{\nu}, \mathbf{G}, T)$ s.t.,*



(a) Graph considered for both $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$.

(b) Expected values $\mu_i(\cdot)$ for $\boldsymbol{\nu}$.

(c) Expected values $\mu_i(\cdot)$ for $\boldsymbol{\nu}'$.

Figure B.1: Instances used in the proof of Theorem 16.

for every policy π, it holds:

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \geqslant \frac{T}{12}.$$

Proof Consider the deterministic rotting scenario, *i.e.*, where $\sigma = 0$. Consider two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ of 3-armed rotting bandit with graph structure as depicted in Figure B.1. The graph which represents the connection of the arms is represented in Figure B.1a. The expected rewards at the different number of triggers n is depicted in Figure B.1b for instance $\boldsymbol{\nu}$ and in Figure B.1c for instance $\boldsymbol{\nu}'$. For both instances, arms 2 and 3 present an expected reward equal to $2/3$ for the first $T/2$ triggers, and then the expected reward becomes 0. On the other hand, the two instances differ in the behavior of the expected reward of arm 1. Indeed, such reward is 1 until we trigger the arm $T/2$ times for instance $\boldsymbol{\nu}$ and for all the T triggers for instance $\boldsymbol{\nu}'$.

We recall that the clairvoyant is aware of both the graph $\mathbf{G}$ and the expected values $\mu_i(n)$, for every $i \in [k]$ and $n \in [T]$. We can easily compute the total reward for the best policy possible π^* for instance $\boldsymbol{\nu}$:

$$J_{\boldsymbol{\nu},\mathbf{G},T}(\pi^*) = \frac{2}{3}T, \tag{B.29}$$

which corresponds to pull arms 2 and 3 only (both for $T/2$ times), and for instance $\boldsymbol{\nu}'$:

$$J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi^*) = T, \tag{B.30}$$

which corresponds to pull always pull arm 1 for T times. We highlight that optimal policy π^* is different for the two instances.

We now need to introduce some additional notations that will be used in the proof. We call $\mathbb{E}_{\boldsymbol{\nu}}\left[N_i^R(n)\right]$ the expected number of pulls for arm i generating reward (*i.e.*, for which the expected reward is different from 0) up to time n for instance $\boldsymbol{\nu}$. We now start by observing that, up to the round $T/2$, the two instances are exactly the same, so every policy π will have the same

behavior in expectation. Given that, we observe that for both the instances we have the same reward, equal to:

$$J_{\boldsymbol{\nu},\mathbf{G},T/2}(\pi) = J_{\boldsymbol{\nu}',\mathbf{G},T/2}(\pi) = \mathbb{E}_{\boldsymbol{\nu}}\left[N_1^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]$$
$$= \frac{T}{2} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right],$$

where the last equality follows from $\mathbb{E}_{\boldsymbol{\nu}}\left[N_1^R\left(\frac{T}{2}\right)\right] + \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] = T/2$. This result is valid for both $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$, as the policy will behave in the same way, and so $\mathbb{E}_{\boldsymbol{\nu}}\left[N_i^R\left(\frac{T}{2}\right)\right] = \mathbb{E}_{\boldsymbol{\nu}'}\left[N_i^R\left(\frac{T}{2}\right)\right]$, for every $i \in [3]$, as the policy on the two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ are not distinguishable the first $T/2$ rounds.

We now have to understand what will happen from $T/2$ to T in the best case possible.

Instance $\boldsymbol{\nu}$ We can easily see how for arm 1 we have terminated the pulls which generate reward, so we have to pull arms 2 and 3. We can now compute the remaining triggers generating reward for arm 2 in the second half of the rounds:

$$\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R(T)\right] - \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] \leqslant \underbrace{\frac{T}{2}}_{\substack{\text{Triggers}\\ \text{initially}\\ \text{available}}} - \underbrace{\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right]}_{\substack{\text{Already}\\ \text{used}}}$$
$$- \underbrace{\left(\frac{T}{2} - \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]\right)}_{\substack{\text{Triggers used}\\ \text{from arm 1}}}$$
$$\leqslant \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]$$

We can do the same reasoning for arm 3 and, for symmetry, we get:

$$\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R(T)\right] - \mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \leqslant \mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right]$$

We now consider a policy using all these triggers, and we compute the expected cumulative reward:

$$J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \leqslant J_{\boldsymbol{\nu},\mathbf{G},T/2}(\pi) + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]$$
$$\leqslant \frac{T}{2} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{2}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]$$
$$\leqslant \frac{T}{2} + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right].$$

Instance $\boldsymbol{\nu}'$ Instead, for instance $\boldsymbol{\nu}'$, we can easily see that the best choice from $T/2$ to T is to always pull arm 1 for all the $T/2$ rounds, receiving a reward of 1 each time. Given that, we have:

$$\begin{aligned} J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi) &\leqslant J_{\boldsymbol{\nu}',\mathbf{G},T/2}(\pi) + \frac{T}{2} \\ &= T - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]. \end{aligned}$$

Regret Moving to the regret, we have for instance $\boldsymbol{\nu}$:

$$\begin{aligned} R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) &= J_{\boldsymbol{\nu},\mathbf{G},T}(\pi^*) - J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \\ &\geqslant \frac{2}{3}T - \frac{T}{2} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] \\ &\geqslant \frac{T}{6} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right], \end{aligned}$$

while for instance $\boldsymbol{\nu}'$:

$$\begin{aligned} R_{\boldsymbol{\nu}',\mathbf{G},T}(\pi) &= J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi^*) - J_{\boldsymbol{\nu}',\mathbf{G},T}(\pi) \\ &\geqslant \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]. \end{aligned}$$

We can now compute a lower bound on the regret:

$$\begin{aligned} R_T(\mathfrak{A}) &= \max\left\{R_{\boldsymbol{\nu},\mathbf{G},T}(\pi), R_{\boldsymbol{\nu}',\mathbf{G},T}(\pi)\right\} \\ &\geqslant \frac{1}{2}\left(R_{\boldsymbol{\nu},\mathbf{G},T}(\pi) + R_{\boldsymbol{\nu}',\mathbf{G},T}(\pi)\right) \\ &= \frac{1}{2}\left(\frac{T}{6} - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] - \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_2^R\left(\frac{T}{2}\right)\right] + \frac{1}{3}\mathbb{E}_{\boldsymbol{\nu}}\left[N_3^R\left(\frac{T}{2}\right)\right]\right) \\ &= \frac{T}{12}. \end{aligned}$$

This proof holds for the specific graph structure we discussed here. However, by joining this result with the one of Lemma 43, we can generalize this result for every non-block-diagonal connectivity matrix. ■

Theorem 14 (Optimal Policy in Rotting GTBs with Block-Diagonal CM)**.** *For any instance $(\boldsymbol{\nu}, \mathbf{G}, T)$ of Rotting GTBs s.t. $\mathbf{G} \in \mathbb{B}_{\tilde{k}}$, the optimal policy $\pi^*_{\boldsymbol{\nu},\mathbf{G},T} \in \arg\max_\pi J_{\boldsymbol{\nu},\mathbf{G},T}(\pi)$ is given by:*

$$\pi^*_{\boldsymbol{\nu},\mathbf{G},T}(t) \in \arg\max_{j\in[k]} \mu_j(\widetilde{N}^*_{j,t}), \qquad \forall t \in [T],$$

where $\widetilde{N}^*_{j,t}$ is the number of times arm j has been triggered by the optimal policy up to time t. Moreover, we have:

$$J^*_{\boldsymbol{\nu},\mathbf{G},T} = \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i \in C_m} \mu_i(n), \tag{3.18}$$

where $N^*_{C_m,T}$ is the number of times the optimal policy pulls an action belonging to clique C_m before T, i.e., $N^*_{C_m,T} = \widetilde{N}^*_{i,T}$, for every $i \in C_m$.

Proof For every Rotting GTB instance, we create an alternative instance which is better, in terms of total cumulative reward, than the original instance. Then we show that playing greedy in the original instance yields the same cumulative reward of the optimal policy from the alternative instance.

For each clique $C_m \in \mathcal{C}_\mathbf{G}$, we substitute the reward function of every arm $i \in C_m$ with $\mu^*_i(n) = \max_{i \in C_m} \mu_i(n)$ for every $n \in [T]$. This way, whenever an action is chosen it is guaranteed to always yield the same reward as any other possible action inside the same clique. We create an alternative instance $(\widetilde{\boldsymbol{\nu}}, \widetilde{\mathbf{G}}, T)$ by collapsing all the actions inside the same clique into a single meta-action, resulting in a $\widetilde{k}$-armed rested rotting bandit problem, where the set of actions corresponds the set of cliques of the original instance. We use Proposition 2 of (Heidari et al., 2016) to get that the optimal policy in the alternative instance is to play, at every round, the action with the highest instantaneous reward. Such policy achieves a total reward, in the alternative instance, of:

$$J^*_{\widetilde{\boldsymbol{\nu}},\widetilde{\mathbf{G}},T} = \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i \in C_m} \mu_i(n),$$

We now show that playing the greedy policy in the original instance yields an equal total cumulative reward. Playing greedily in the original instance we get:

$$\begin{aligned}
J^*_{\boldsymbol{\nu},\mathbf{G},T} &= \sum_{t=1}^{T} \max_{i \in [k]} \mu_i(\widetilde{N}^*_{i,t}) \\
&= \sum_{C_m \mathcal{C}_\mathbf{G}} \sum_{t=1}^{T} \mathbb{1}_{\{I^*_t \in C_m\}} \max_{i \in [k]} \mu_i(\widetilde{N}^*_{i,t}) \\
&= \sum_{C_m \mathcal{C}_\mathbf{G}} \sum_{t=1}^{T} \mathbb{1}_{\{I^*_t \in C_m\}} \max_{i \in C_m} \mu_i(\widetilde{N}^*_{i,t}) \\
&= \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i \in C_m} \mu_i(n) \\
&= J^*_{\widetilde{\boldsymbol{\nu}},\widetilde{\mathbf{G}},T} \geqslant J_{\boldsymbol{\nu},\mathbf{G},T}(\pi) \quad \forall \pi.
\end{aligned}$$

The performance of the optimal policy in the alternative instance is matched by the greedy policy played in the original instance. The proof is concluded by observing that the optimal total cumulative reward of the alternative instance cannot be lower than the total reward of any policy π in the original instance, since the alternative instance has pointwise higher reward functions for every action. ■

B.2.1. Upper Bounding the Regret of RAW-UCB

We start by defining the expectation version of the estimator defined in Equation (3.19) as $\bar{\mu}_i^h(t) := \frac{1}{h}\sum_{s=1}^{t-1} \mathbb{1}_{\{I_t=i \,\wedge\, N_{i,s}>N_{i,t-1}-h\}}\mu_i(\widetilde{N}_{i,s})$. Before moving on, we recall the following result, which also introduces the notion of *good event* ξ_t^α.

Proposition B.1 (Bound on the Probability of Bad Event, Seznec et al. 2020)**.** *Let $\delta_t = 2t^{-\alpha}$, and*

$$\xi_t^\alpha := \left\{\forall i \in [k],\ \forall n \leqslant t-1,\ \forall h \leqslant n,\ |\widehat{\mu}_i^h(t) - \bar{\mu}_i^h(t)| \leqslant c(h,\delta_t)\right\},$$

for $c(h,\delta_t) := \sqrt{2\sigma^2 \log(2\delta_t^{-1})/h}$. Then

$$\mathbb{P}\left(\bar{\xi_t^\alpha}\right) \leqslant Kt^{2-\alpha}. \tag{B.31}$$

Lemma 44 (Overestimation under the Good Event)**.** *Under ξ_t^α, if action I_t is selected by Algorithm 8, for every $h \in [N_{i,t-1}]$ we have:*

$$\bar{\mu}_{I_t}^h \geqslant \max_{i\in[k]} \mu_i(\widetilde{N}_{i,t-1}^\pi) - 2c(h,\delta_t), \tag{B.32}$$

where $\widetilde{N}_{i,t-1}^\pi$ is the number of triggers of action i provoked by playing with Algorithm 8 up until time t.

Proof This proof is adapted from the one of Lemma 1 of (Seznec et al., 2020). Let $h_{i,t}^{\min} \in \arg\min_{h\leqslant N_{i,t-1}} \widehat{\mu}_i^h(t) + c(h,\delta_t)$.

Let $i_t^\pi \in \arg\max_{i\in[k]} \mu_i(\widetilde{N}_{i,t-1}^\pi)$ be the best available action at time t. From the rotting assumption, we know that:

$$\max_{i\in[k]} \mu_i(\widetilde{N}_{i,t-1}^\pi) = \mu_{i_t^\pi}(\widetilde{N}_{i,t-1}^\pi) \leqslant \bar{\mu}_{i_t^\pi}^1(t) \leqslant \ldots \leqslant \bar{\mu}_{i_t^\pi}^{h_{i_t^\pi,t}^{\min}}(t).$$

Under ξ_t^α, we have:

$$\bar{\mu}_{i_t^\pi}^{h_{i_t^\pi,t}^{\min}}(t) \leqslant \widehat{\mu}_{I_t}^{h_{I_t,t}^{\min}}(t) + c(h_{I_t,t}^{\min}, \delta_t).$$

We now use the definition of $h_{I_t,t}^{\min}$:

$$\widehat{\mu}_{I_t}^{h_{I_t,t}^{\min}}(t) + c(h_{I_t,t}^{\min}, \delta_t) \leqslant \widehat{\mu}_{I_t}^{h}(t) + c(h, \delta_t).$$

Again, we use ξ_t^α:

$$\widehat{\mu}_{I_t}^{h}(t) + c(h, \delta_t) \leqslant \bar{\mu}_{I_t}^{h}(t) + 2c(h, \delta_t).$$

Putting all together, we obtain the statement. ■

Theorem 15 (RAW-UCB Regret in Rotting GTBs with Block-Diagonal CM)**.** *Let* $(\boldsymbol{\nu}, \mathbf{G}, T)$ *be an instance of the Rotting GTBs, where* $\mathbf{G} \in \mathbf{B}_{\tilde{k}}$*. Let* $\delta_t = t^{-\alpha}$ *for* $\alpha \geqslant 5$*. Then,* RAW-UCB *suffers an expected regret bounded as:*

$$R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB}) \leqslant \widetilde{\mathcal{O}}\Bigg(\underbrace{k\left(\sigma\sqrt{\log T} + V_{\boldsymbol{\nu}}(T)\right)}_{\text{(A) Variance Contribution}} + \underbrace{L\sum_{C_m \in \mathcal{C}_{\mathbf{G}}} |C_m|^2 + kL + \sigma \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(\sqrt{\frac{|C_m|}{k}T}\right)}_{\text{(B) Rested Contribution}} + \underbrace{(\alpha\sigma)^{\frac{2}{3}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(V_T^\pi \frac{|C_m|}{k} T^2\right)^{\frac{1}{3}}}_{\text{(C) Restless Contribution}} \Bigg).$$

Proof Let us proceed to decompose the regret:

$$\begin{aligned}
R_{\boldsymbol{\nu},\mathbf{G},T}(\texttt{RAW-UCB}) &= \sum_{t=1}^{T} \left(\mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \mu_{I_t}(\widetilde{N}_{I_t,t}^\pi)\right) \\
&= \sum_{t=1}^{T} \left(\mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \mu_{I_t}(\widetilde{N}_{I_t,t}^\pi) \pm \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi)\right) \\
&= \underbrace{\sum_{t=1}^{T} (\mu_{i_t^*}(\widetilde{N}_{i_t^*,t}^*) - \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi))}_{\text{(b)}} + \underbrace{\sum_{t=1}^{T} (\max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi) - \mu_{I_t}(\widetilde{N}_{I_t,t}^\pi))}_{\text{(c)}}
\end{aligned}$$

Before bounding the two terms, we observe the following:

$$\sum_{t=1}^{T} \max_{i \in C_{I_t}} \mu_i(\widetilde{N}_{I_t,t}^\pi) = \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N_{C_m,T}^\pi} \max_{i \in C} \mu_i(n). \tag{B.33}$$

Equation (B.33) is a consequence of Equation (3.18) (Theorem 14), when applied to the restless bandit problems obtained by each clique when considered alone. We have for (c):

$$\begin{aligned}
\text{(c)} &= \sum_{t=1}^{T} \max_{i \in C_{I_t}} \left(\mu_i(\widetilde{N}^{\pi}_{I_t,t}) - \mu_{I_t}(\widetilde{N}^{\pi}_{I_t,t}) \right) \\
&\overset{\text{Eq. (B.33)}}{=} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^{\pi}_{C_m,T}} \left(\max_{i \in C} \mu_i(n) - \mu_{I_{t_{C,n}}}(n) \right) \\
&\overset{(\lozenge)}{\leqslant} 6kV_{\boldsymbol{\nu}}(T) + 4(8\alpha\sigma)^{\frac{2}{3}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(V_{\boldsymbol{\nu}}(T)|C_m|(N^{\pi}_{C_m,T})^2 \log T \right)^{\frac{1}{3}} + \\
&\quad + 2(2\sqrt{2}\alpha\sigma)^{\frac{1}{3}} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \left(V_{\boldsymbol{\nu}}(T)^2 |C_m|^2 N^{\pi}_{C_m,T} \sqrt{\log T} \right)^{\frac{1}{3}}
\end{aligned}$$

The last inequality is obtained by observing that, fixing the number of times a pull is selected, we have a nested restless bandit problem having as the time horizon the number of times the clique is pulled $N^{\pi}_{C_m,T}$. `RAW-UCB` plays greedily in each clique independently. Thus, when an action belonging to clique C is selected, it is the same action that an instance of `RAW-UCB` would have played in a restless rotting bandit composed only by the actions belonging to C.[1] In the step marked with $(\lozenge)$, this equivalence allows us to bound (c) with the summation of regret bounds of the algorithm for smaller restless bandits defined for the cliques, by using Theorem 1 from (Seznec et al., 2020) and bounding $\log N^{\pi}_{C_m,T} \leqslant \log T$ and $V_{\boldsymbol{\nu}}(N^{\pi}_{C_m,T}) \leqslant V_{\boldsymbol{\nu}}(T)$ for every $C_m \in \mathcal{C}_{\mathbf{G}}$. The last term is dominated by the other two in every quantity, and is thus omitted in the final bound.

We now focus on (b):

$$\begin{aligned}
\text{(b)} &= \sum_{t=1}^{T} (\mu_{i^*_t}(\widetilde{N}^{*}_{i^*_t,t}) - \max_{i \in C_{I_t}} \mu_i(\widetilde{N}^{\pi}_{I_t,t})) \\
&\overset{\text{(B.32)}}{=} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^{*}_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{t=1}^{T} \max_{i \in C_{I_t}} \mu_i(\widetilde{N}^{\pi}_{I_t,t}) \\
&\overset{\text{(B.33)}}{=} \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^{*}_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} \sum_{n=1}^{N^{\pi}_{C_m,T}} \max_{i \in C} \mu_i(n)
\end{aligned}$$

The term (b) only depends on the difference between the allocation of pulls among the cliques between the optimal policy and the algorithm's policy. Thus, it makes sense to split the cliques into two sets, namely OP and UP: the first will contain the OverPulled cliques, the second the

[1] The UCB is different since the clique-specific instance of `RAW-UCB` would have used *internal times* instead of the external time t, however, the order is preserved and the decision is the same.

UnderPulled cliques, which are cliques pulled by RAW-UCB more than the optimal policy and the cliques pulled less, respectively.

$$\sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^*_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{C_m \in \mathcal{C}_\mathbf{G}} \sum_{n=1}^{N^\pi_{C_m,T}} \max_{i \in C} \mu_i(n)$$
$$= \sum_{C_m \in \text{UP}} \sum_{n=N^\pi_{C_m,T}+1}^{N^*_{C_m,T}} \max_{i \in C} \mu_i(n) - \sum_{C_m \in \text{OP}} \sum_{n=N^*_{C_m,T}+1}^{N^\pi_{C_m,T}} \max_{i \in C} \mu_i(n)$$

We now introduce the auxiliary quantity $\mu^+_T(\pi) \coloneqq \max_{i \in [k]} \mu_i(\widetilde{N}^\pi_{i,T})$. We also observe that the two terms in the RHS have the same number of addends, since the number of overpulls must be equal to the number of underpulls. Finally, we define $h_{C,T}$ as the number of overpulls of clique C.

$$\sum_{C_m \in \text{UP}} \sum_{n=N^\pi_{C_m,T}}^{N^*_{C_m,T}-1} \max_{i \in C} \mu_i(n) - \sum_{C_m \in \text{OP}} \sum_{n=N^*_{C_m,T}}^{N^\pi_{C_m,T}-1} \max_{i \in C} \mu_i(n)$$
$$\leqslant \sum_{C_m \in \text{UP}} \sum_{n=N^\pi_{C_m,T}}^{N^*_{C_m,T}-1} \mu^+_T(\pi) - \sum_{C_m \in \text{OP}} \sum_{n=N^*_{C_m,T}}^{N^\pi_{C_m,T}-1} \max_{i \in C} \mu_i(n)$$
$$= \sum_{C_m \in \text{OP}} \sum_{n=N^*_{C_m,T}}^{N^\pi_{C_m,T}-1} (\mu^+_T(\pi) - \max_{i \in C} \mu_i(n))$$
$$= \sum_{C_m \in \text{OP}} \sum_{h=0}^{h_{C,T}-1} (\mu^+_T(\pi) - \max_{i \in C} \mu_i(N^*_{C_m,T} + h)).$$

We can now decompose the last summation by the means of events $\{\xi^\alpha_t\}_t$:

$$(\text{b}_\xi) \leqslant \sum_{C_m \in \text{OP}} \sum_{h=0}^{h_{C,T}-1} \mathbb{1}\{\xi^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}}\}(\mu^+_T(\pi) - \max_{i \in C} \mu_i(N^*_{C_m,T} + h))$$
$$\leqslant \sum_{C_m \in \text{OP}^\xi} \sum_{h=0}^{h^\xi_{C,T}} (\mu^+_T(\pi) - \max_{i \in C} \mu_i(N^*_{C_m,T} + h)),$$

where $h^\xi_{C,T} \coloneqq \max\{h \leqslant h_{C,T} : \xi_{t^\pi_{C,N^*_{C_m,T}+h}}\}$ is the largest number of overpulls a clique undergoes before time $t^\pi_{C,N^*_{C_m,T}+h} \leqslant T$ under the events ξ^α_t, and $\text{OP}^\xi \coloneqq \{C_m \in \text{OP} : h^\xi_{C,T} \geqslant 1\}$. We call, for short, $\widetilde{t}_{C,h}$ the time at which clique C is overpulled for the h-th time i.e., $t^\pi_{C,N^*_{C_m,T}+h}$,

and observe that

$$
\begin{aligned}
&\sum_{h=0}^{h^{\xi}_{C,T}} \max_{i\in C}\mu_i(N^*_{C_m,T}+h)\\
&\qquad= \sum_{h=0}^{h^{\xi}_{C,T}} \mathbb{1}\{h \neq h_{C,t_{j,N^{\pi}_{j,T}}}\ \forall j\in[k]\} \max_{i\in C}\mu_i(N^*_{C_m,T}+h) + \sum_{j\in C}\max_{i\in C}\mu_i(N^*_{C_m,T}+h_{C,t_{j,N^{\pi}_{j,T}}})\\
&\qquad= \sum_{i\in C}\sum_{h=0}^{h^{\xi}_{i,T}-1}\mu_i(N^{\pi}_{C,\tilde{t}_{C,h}}) + \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})\\
&\qquad\overset{(3.19)}{=} \sum_{i\in C}(h^{\xi}_{i,T}-1)\bar{\mu}_i^{h^{\xi}_{i,T}-1}(\tilde{t}_{C,h^{\xi}_{i,T}}) + \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})\\
&\qquad\overset{(B.32)}{\geqslant} \sum_{i\in C}(h^{\xi}_{i,T}-1)\left(\max_{i\in[k]}\mu_i(\widetilde{N}^{\pi}_{i,T}) - 2c(h^{\xi}_{i,T}-1,\delta_{\tilde{t}_{C,h^{\xi}_{i,T}}})\right) + \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})\\
&\qquad\geqslant (h^{\xi}_{C,T}-|C_m|)\max_{i\in[k]}\mu_i(\widetilde{N}^{\pi}_{i,T}) - 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T) + \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}})\\
&\qquad= (h^{\xi}_{C,T}-|C_m|)\mu^{+}_T(\pi) - 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T) + \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}}).
\end{aligned}
$$

Plugging this observation into the previous, we get:

$$
\begin{aligned}
(\mathrm{b}_\xi) &\leqslant \sum_{C_m\in\mathrm{OP}^{\xi}}\left(|C_m|\mu^{+}_T(\pi) - \sum_{j\in C}\max_{i\in C}\mu_i(\widetilde{N}^{\pi}_{j,t_{j,N^{\pi}_{j,T}}}) + 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T)\right)\\
&= \sum_{C_m\in\mathrm{OP}^{\xi}}\left(\sum_{i\in C}(\mu^{+}_T(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}})) + 2\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T)\right)\\
&\overset{(\star)}{\leqslant} 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in\mathrm{OP}^{\xi}}\left(\sum_{i\in C}(h^{\xi}_{i,T}-1)c(h^{\xi}_{i,T}-1,\delta_T)\right)\\
&\leqslant 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in\mathrm{OP}^{\xi}}\left(\sigma\sum_{i\in C}\sqrt{(h^{\xi}_{i,T}-1)\log T}\right)\\
&\leqslant 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in\mathrm{OP}}\left(\sigma\sqrt{\log T}\sum_{i\in C}\sqrt{(h_{i,T}-1)}\right)\\
&\overset{(J)}{\leqslant} 2k\sigma\sqrt{\log T} + L\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}|C_m|^2 + 2\sum_{C_m\in\mathcal{C}_{\mathbf{G}}}\left(\sigma\sqrt{|C_m|N^{\pi}_{C_m,T}\log T}\right).
\end{aligned}
$$

The step marked with $(\star)$ is justified by the following considerations. Let $i\in C$, we shorten the notation for the time at which clique C is triggered for the $(\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}}-m)$-th time as $t_{i,-m} := t_{C,\widetilde{N}^{\pi}_{i,t_{i,N^{\pi}_{i,T}}}-m}$. In other words, after this time the clique C is only chosen m times before

the action $i \in C$ is pulled for the last time. Consider the $|C_m|$ times the clique C is chosen before pulling i for the last time: then, due to the pigeonhole principle, at least one action belonging to the clique should appear at least two times before the last pull. Without loss of generality, we assume that only one action appears exactly two times, and call the first appearance time $t_{i,-m}$ and the second $t_{i,-m'}$ (note that $m' \leqslant m \leqslant |C_m|$). We now observe that:

$$\begin{aligned}\sum_{i\in C}\left(\mu_T^+(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^\pi_{i,t_{i,N^\pi_{i,T}}})\right) &= \sum_{i\in C}\left(\mu_T^+(\pi) - \max_{j\in C}\left\{\mu_j(\widetilde{N}^\pi_{i,t_{i,N^\pi_{i,T}}}) \pm \mu_j(\widetilde{N}^\pi_{i,t_{i,N^\pi_{i,T}}} - m)\right\}\right)\\ &\leqslant \sum_{i\in C}\left(\mu_T^+(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^\pi_{i,t_{i,N^\pi_{i,T}}} - m) + mL\right)\\ &\leqslant \sum_{i\in C}\left(\mu_T^+(\pi) - \max_{j\in C}\mu_j(\widetilde{N}^\pi_{i,t_{i,N^\pi_{i,T}}} - m)\right) + |C_m|^2 L,\end{aligned}$$

We can now prove the step $(\star)$ by bounding

$$\begin{aligned}\sum_{i\in C}\max_{j\in C}\mu_j(\widetilde{N}^\pi_{i,t_{i,N^\pi_{i,T}}} - m) &\geqslant \sum_{i\in C}\mu_{I_{t_{i,-m}}}(\widetilde{N}^\pi_{i,t_{i,N^\pi_{i,T}}} - m)\\ &= \sum_{i\in C}\bar{\mu}^1_{I_{t_{i,-m}}}(t_{i,-m'})\\ &\overset{(B.32)}{\geqslant} \sum_{i\in C}\left(\max_{j\in[k]}\mu_j(\widetilde{N}^\pi_{j,t_{i,-m'}}) - 2c(1,\delta_{t_{i,-m'}})\right)\\ &\overset{(\dagger)}{\geqslant} \sum_{i\in C}\left(\mu_T^+(\pi) - 2c(1,\delta_T)\right),\end{aligned}$$

where $(\dagger)$ is a consequence of $\widetilde{N}^\pi_{j,t_{i,-m'}} \leqslant \widetilde{N}^\pi_{j,T}$ for every $j \in [k]$.

Finally, in the step marked with (J) we used Jensen inequality to find the worst allocation of overpull among the actions in the same clique, which is the uniform one, i.e., $h_{i,T} \leqslant N^\pi_{C_m,T}/|C_m|$.

To conclude the proof, we need to find out what happens under $\bar{\xi}_t^{\,\alpha}$:

$$\begin{aligned}(\mathrm{b}_{\bar{\xi}}) &\leqslant \sum_{C_m\in\mathrm{OP}}\sum_{h=0}^{h_{C,T}-1}\mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}}\}(\mu_T^+(\pi) - \max_{i\in C}\mu_i(N^*_{C_m,T}+h))\\ &= \sum_{C_m\in\mathrm{OP}}\sum_{h=0}^{h_{C,T}-1}\sum_{t=1}^{T}\mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}} \wedge t^\pi_{C,N^*_{C_m,T}+h} = t\}(\mu_T^+(\pi) - \max_{i\in C}\mu_i(N^*_{C_m,T}+h))\\ &\overset{(*)}{\leqslant} \sum_{t=1}^{T}\mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}}\}Lt\left(\sum_{C_m\in\mathrm{OP}}\sum_{h=0}^{h_{C,T}-1}\mathbb{1}\{t^\pi_{C,N^*_{C_m,T}+h} = t\}\right)\\ &\leqslant \sum_{t=1}^{T}\mathbb{1}\{\bar{\xi}^\alpha_{t^\pi_{C,N^*_{C_m,T}+h}}\}Lt.\end{aligned}$$

In the step marked with $(\star)$, we use the fact that at time t overpulling a clique can yield at most Lt regret. The last step is a consequence that for each round t we can have at most 1 overpull. We conclude the bound by using Proposition B.1:

$$\mathbb{E}[(\mathrm{b}_{\bar{\xi}})] \leqslant \sum_{t=1}^{T} \mathbb{P}\left(\bar{\xi}_t^{\alpha}\right) Lt \overset{\text{(B.31)}}{\leqslant} \sum_{t=1}^{T} kLt^{3-\alpha} \overset{(\alpha \geqslant 5)}{\leqslant} 2kL.$$

We then observe that, given an arbitrary concave function g, we have

$$\sum_{C_m \in \mathcal{C}_{\mathbf{G}}} g(|C_m| N_{C_m,T}^{\pi}) \leqslant \sum_{C_m \in \mathcal{C}_{\mathbf{G}}} g\left(\frac{|C_m|}{k} T\right).$$

This can be applied to component (b) with $g(\cdot) = \sqrt{\cdot}$ and to component (c) with $g(\cdot) = (\cdot)^{\frac{2}{3}}$. The statement of the theorem can be obtained by summing up all the components, i.e.,

$$\mathbb{E}[R_T(\pi)] \leqslant \mathbb{E}\left[(\mathrm{b}_{\xi}) + (\mathrm{b}_{\bar{\xi}}) + (\mathrm{c})\right].$$

■

C | Rising Bandits

C.1. Lower Bounds

In this appendix, we provide the proofs of the results presented in Section 3.3.4 in the main paper.

C.1.1. General Recipe for the Lower Bound

The goal of this section is to prove Lemma 17. Remember that we work under the assumption of Bernoulli-distributed rewards. The result is obtained through techniques from the adversarial literature in which the instance is also affected by randomness. Thus, we define two probability distributions over $\{0, \dots, k\}^{\mathbb{N}_{\geqslant 1}}$, which induce probability distributions over the instances in $\mathcal{E}_{\overline{\mu},\widetilde{\mu}}$. In particular, let $\overline{\xi}, \widetilde{\xi} \in D(\{0, \dots, k\})$ defined as:

$$\overline{\xi}(\{o\}) := \begin{cases} 0 & \text{if } o \in [k], \\ 1 & \text{if } o = 0 \end{cases}$$

$$\widetilde{\xi}(\{o\}) := \begin{cases} \frac{1}{k} & \text{if } o \in [k], \\ 0 & \text{if } o = 0 \end{cases}$$

for $o \in \{0, \dots, k\}$. We can extend $\overline{\xi}$ and $\widetilde{\xi}$ to probability distributions over $\{0, \dots, k\}^{\mathbb{N}_{\geqslant 1}}$ via infinite product (see example 1.63 of Klenke 2020):

$$\overline{\tau}_w := \left(\otimes_{l=1}^{w-1} \widetilde{\xi}\right) \otimes \left(\otimes_{l=w}^{+\infty} \overline{\xi}\right) \text{ for all } w \in \mathbb{N}_{\geqslant 1},$$

$$\widetilde{\tau} := \otimes_{w \in \mathbb{N}_{\geqslant 1}} \widetilde{\xi}.$$

$\widetilde{\tau}$ models a random instance in which, in each window, we choose independently and uniformly one arm whose expected reward follows the modified trend, while the expected rewards of all the other arms follow the base trend. $\overline{\tau}_w$ instead models a random instance which behaves like $\widetilde{\tau}$ up to window $w \in \mathbb{N}_{\geqslant 1}$ (excluded); from window w onward all arms

follow the base trend. For technical reasons which will be clear in what follows, we need to build a probability space in which the randomness over the instance and the randomness over the rewards are unlinked. Observe that with the current construction this is not the case. Indeed, $\boldsymbol{X}$ is sampled from $\boldsymbol{\nu_o}$, but $\boldsymbol{o}$ is also a random element. To this end, let $\boldsymbol{s} = (s_{i,t})_{i\in[k],t\in\mathbb{N}_{\geqslant 1}} \sim \lambda := \otimes_{i\in[k],t\in\mathbb{N}_{\geqslant 1}} \mathrm{Unif}(0,1)$ where $\mathrm{Unif}(0,1)$ is the uniform distribution with support $[0,1]$. Then, we can redefine $X_{i,t}(\boldsymbol{o},\boldsymbol{s}) = \mathbf{1}[s_{i,t} \leqslant \mu_{\boldsymbol{o},i}(t)]$ where $\mu_{\boldsymbol{o},i}(t)$ is defined in analogy to Equation (3.21). In this way, we moved the dependency from $\boldsymbol{o}$ inside the definition of the random variables, preserving their distributions. For consistency with the notation, we introduce the random variables $\boldsymbol{O} = (O_w)_{w\in\mathbb{N}_{\geqslant 1}}$ where $O_w(\boldsymbol{o}) = o_w$. The probability distributions that we just defined, induce probability density functions over finite reward sequences taking into account the randomness both in the instance and in the rewards. In particular, let

$$\overline{p}_w(r_1,\dots,r_T) := \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\overline{\tau}_w\\ \boldsymbol{s}\sim\lambda}}[R_1 = r_1,\dots,R_T = r_T],$$

$$\widetilde{p}_{w,i}(r_1,\dots,r_T) := \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[R_1 = r_1,\dots,R_T = r_T \mid O_w = i]$$

for $w \in \mathbb{N}_{\geqslant 1}$, $i \in [k]$, $r_1,\dots,r_T \in \{0,1\}$. We use $\overline{p}_w$ and $\widetilde{p}_{w,i}$ to denote also all the conditional and marginal distributions; disambiguation happens through the arguments, *e.g.*, $\overline{p}_w(r_{s_w} \mid r_1,\dots,r_{s_w-1})$.

To obtain the result, we use the following tools from information theory (Cover and Thomas, 2006).

Definition 5 (L^1 Distance of Two Discrete Probability Density Functions)**.** *Let p, q be two discrete probability density functions defined over the finite set $\mathcal{X}$, we define their L^1 distance as:*

$$\|p - q\|_1 := \sum_{x\in\mathcal{X}} |p(x) - q(x)|.$$

Definition 6 (Kullback-Leibler Divergence of Two Discrete Probability Density Functions)**.** *Let p, q be two discrete probability density functions defined over the finite set $\mathcal{X}$, we define their Kullback-Leibler divergence as:*

$$\mathrm{D}_{\mathrm{KL}}(p\|q) := \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{p(x)}{q(x)}\right).$$

We use $\mathrm{D}_{\mathrm{KL}}(x_1\|x_2)$ with $x_1, x_2 \in [0,1]$ to denote the Kullback-Leibler divergence of two Bernoulli p.d.f.s with corresponding expected values x_1 and x_2.

We now state and prove a generalization of Lemma A.1 in (Auer et al., 2002b) which we then

use to derive Lemma 17.

Lemma 45. *Let $w \in [w(T)]$, $i \in [k]$, $f : \{0,1\}^{\min\{e_w,T\}} \to [0,M]$ with $M \geqslant 0$. Then:*

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[f(R_1,\dots,R_{\min\{e_w,T\}}) \mid O_w = i] - \mathop{\mathbb{E}}_{\substack{\boldsymbol{o}\sim\overline{\tau}_w\\ \boldsymbol{s}\sim\lambda}}[f(R_1,\dots,R_{\min\{e_w,T\}})]$$
$$\leqslant \frac{M}{2}\sqrt{2\ln(2)\sum_{t=s_w}^{\min\{e_w,T\}} \mathrm{D}_{\mathrm{KL}}(\overline{\mu}_w(t-s_w+1)\|\widetilde{\mu}_w(t-s_w+1))\mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\overline{\tau}_w\\ \boldsymbol{s}\sim\lambda}}[I_t = i]}. \quad \text{(C.1)}$$

Proof To simplify the notation, let $t_1 := s_w, t_2 := \min\{e_w, T\}$. The lhs of Equation (C.1) can be written as:

$$\begin{aligned}
&\sum_{r_1,\dots,r_{t_2}\in\{0,1\}} f(r_1,\dots,r_{t_2})\,(\widetilde{p}_{w,i}(r_1,\dots,r_{t_2}) - \overline{p}_w(r_1,\dots,r_{t_2}))\\
&\leqslant M \sum_{\substack{r_1,\dots,r_{t_2}\in\{0,1\}\\ \text{s.t. } \widetilde{p}_{w,i}(r_1,\dots,r_{t_2})\geqslant\overline{p}_w(r_1,\dots,r_{t_2})}} (\widetilde{p}_{w,i}(r_1,\dots,r_{t_2}) - \overline{p}_w(r_1,\dots,r_{t_2}))\\
&= \frac{M}{2}\|\overline{p}_w(r_1,\dots,r_{t_2}) - \widetilde{p}_{w,i}(r_1,\dots,r_{t_2})\|_1, \qquad \text{(C.2)}
\end{aligned}$$

where line (C.2) can be found in (Chapter 11, Cover and Thomas, 2006). Again, from (Lemma 11.6.1 Cover and Thomas, 2006), we have that:

$$\|\overline{p}_w(r_1,\dots,r_{t_2}) - \widetilde{p}_{w,i}(r_1,\dots,r_{t_2})\|_1^2 \leqslant 2\ln(2)\mathrm{D}_{\mathrm{KL}}(\overline{p}_w(r_1,\dots,r_{t_2})\|\widetilde{p}_{w,i}(r_1,\dots,r_{t_2})).$$

From the chain rule of entropy:

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\overline{p}_w(r_1,\dots,r_{t_2})\|\widetilde{p}_{w,i}(r_1,\dots,r_{t_2})) = &\underbrace{\sum_{t=t_1}^{t_2} \mathrm{D}_{\mathrm{KL}}(\overline{p}_w(r_t \mid r_1,\dots,r_{t-1})\|\widetilde{p}_{w,i}(r_1 \mid r_1,\dots,r_{t-1}))}_{\text{(a)}}\\
&+ \underbrace{\mathrm{D}_{\mathrm{KL}}(\overline{p}_w(r_1,\dots,r_{t_1-1})\|\widetilde{p}_{w,i}(r_1,\dots,r_{t_1-1}))}_{\text{(b)}}.
\end{aligned}$$

Because of how $\overline{\tau}_w$ and $\widetilde{\tau}$ are defined, we have that:

$$\widetilde{p}_{w,i}(r_1,\dots,r_{t_1-1}) = \overline{p}_w(r_1,\dots,r_{t_1-1}) \quad \text{for all} \quad r_1,\dots,r_{t_1-1} \in \{0,1\}$$

and thus term (b) is 0 because of the properties of $\mathrm{D}_{\mathrm{KL}}(\cdot\|\cdot)$. To deal with term (a) we need to work on the expressions of $\widetilde{p}_{w,i}(r_t \mid r_1,\dots,r_{t-1})$ and $\overline{p}_w(r_t \mid r_1,\dots,r_{t-1})$ for $t \in \{t_1,\dots,t_2\}$. First of all observe that the arm that the agent pulls at round t is fully determined by the past

sequence of observed rewards $r_1, \dots, r_{t-1}$ since the policy π is deterministic. As remarked in Section 3.1.2, we denote it through $\pi(t)$, omitting the dependence on $r_1, \dots, r_{t-1}$. Now:[1]

$$\begin{aligned}
\widetilde{p}_{w,i}(r_1, \dots, r_t) &= \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau} \\ \boldsymbol{s}\sim\lambda}}[R_1 = r_1, \dots, R_t = r_t \mid O_w = i] \\
&= \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau} \\ \boldsymbol{s}\sim\lambda}}[X_{\pi(1),1} = r_1, \dots, X_{\pi(t),t} = r_t \mid O_w = i] \\
&= \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau} \\ \boldsymbol{s}\sim\lambda}}[X_{\pi(1),1} = r_1, \dots, X_{\pi(t-1),t-1} = r_{t-1} \mid O_w = i] \\
&\cdot (\mathbf{1}[\pi(t) = i] \operatorname{Be}(r_t \mid \widetilde{\mu}_w(t - s_w + 1)) + \mathbf{1}[\pi(t) \neq i] \operatorname{Be}(r_t \mid \overline{\mu}_w(t - s_w + 1))) \\
&= \widetilde{p}_{w,i}(r_1, \dots, r_{t-1})(\mathbf{1}[\pi(t) = i] \operatorname{Be}(r_t \mid \widetilde{\mu}_w(t - s_w + 1)) \\
&+ \mathbf{1}[\pi(t) \neq i] \operatorname{Be}(r_t \mid \overline{\mu}_w(t - s_w + 1))), \qquad \text{(C.3)}
\end{aligned}$$

where line (C.3) follows from the fact that, under the event $O_w = i$, $X_{\pi(t),t}$ is independent from $X_{\pi(1),1}, \dots, X_{\pi(t-1),t-1}$ and has expected value $\widetilde{\mu}_w(t - s_w + 1)$ if $\pi(t) = i$, $\overline{\mu}_w(t - s_w + 1)$ otherwise. Thus we conclude:

$$\widetilde{p}_{w,i}(r_t \mid r_1, \dots, r_{t-1}) = \mathbf{1}[\pi(t) = i] \operatorname{Be}(r_t \mid \widetilde{\mu}_w(t - s_w + 1)) + \mathbf{1}[\pi(t) \neq i] \operatorname{Be}(r_t \mid \overline{\mu}_w(t - s_w + 1)).$$

From analogous calculations it is possible to derive:

$$\overline{p}_w(r_t \mid r_1, \dots, r_{t-1}) = \operatorname{Be}(r_t \mid \overline{\mu}_w(t - s_w + 1)).$$

[1] With slight abuse of notation, we will use the symbol $\operatorname{Be}(x)$ also to denote the p.d.f. of a Bernoulli distribution of parameter $x \in [0, 1]$.

Thanks to the last results and the definition of $\mathrm{D}_{\mathrm{KL}}(\cdot\|\cdot)$:

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\overline{p}_w(r_1,\dots,r_{t_2})\|\widetilde{p}_{w,i}(r_1,\dots,r_{t_2})) = \sum_{t=t_1}^{t_2} \sum_{r_1,\dots,r_t\in\{0,1\}} \overline{p}_w(r_1,\dots,r_t) \\
&\cdot \log_2\left(\frac{\mathrm{Be}(r_t \mid \overline{\mu}_w(t-s_w+1))}{\mathbf{1}[\pi(t)=i]\,\mathrm{Be}(r_t \mid \widetilde{\mu}_w(t-s_w+1)) + \mathbf{1}[\pi(t)\neq i]\,\mathrm{Be}(r_t \mid \overline{\mu}_w(t-s_w+1))}\right) \\
&= \sum_{t=t_1}^{t_2} \sum_{r_1,\dots,r_{t-1}\in\{0,1\}} \overline{p}_w(r_1,\dots,r_{t-1})\mathbf{1}[\pi(t)=i] \sum_{r_t\in\{0,1\}} \mathrm{Be}(r_t \mid \overline{\mu}_w(t-s_w+1)) \\
&\cdot \log_2\left(\frac{\mathrm{Be}(r_t \mid \overline{\mu}_w(t-s_w+1))}{\mathrm{Be}(r_t \mid \widetilde{\mu}_w(t-s_w+1))}\right) \\
&= \sum_{t=t_1}^{t_2} \mathrm{D}_{\mathrm{KL}}(\overline{\mu}_w(t-s_w+1)\|\widetilde{\mu}_w(t-s_w+1)) \sum_{r_1,\dots,r_{t-1}\in\{0,1\}} \overline{p}_w(r_1,\dots,r_{t-1})\mathbf{1}[\pi(t)=i] \\
&= \sum_{t=s_w}^{\min\{e_w,T\}} \mathrm{D}_{\mathrm{KL}}(\overline{\mu}_w(t-s_w+1)\|\widetilde{\mu}_w(t-s_w+1)) \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\overline{\tau}_w \\ \boldsymbol{s}\sim\lambda}} [I_t=i].
\end{aligned}
$$

The lemma follows by chaining the results. ■

We are ready to prove Lemma 17.

Lemma 17 (General Lower Bound)**.** *Under the assumption that $\overline{\mu}_w(t) \leqslant \widetilde{\mu}_w(t)$ for all $w \in \mathbb{N}_{\geqslant 1}$, $t \in [D_w]$, for any deterministic policy π and learning horizon $T \in \mathbb{N}_{\geqslant 1}$, assuming Bernoulli-distributed rewards, it holds that:*

$$
\sup_{\boldsymbol{\nu}\in\mathcal{E}_{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}}}} R_{\boldsymbol{\nu}}(\pi,T) \geqslant \sum_{w=1}^{w(T)} \left(1 - \frac{1}{k} - \frac{1}{\sqrt{2k}}\sqrt{\ln(2) D_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}}\right) A_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}, \tag{3.22}
$$

where:

$$
\begin{aligned}
D_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T} &:= \sum_{t=s_w}^{\min\{e_w,T\}} \mathrm{D}_{\mathrm{KL}}(\overline{\mu}_w(t-s_w+1)\|\widetilde{\mu}_w(t-s_w+1)), \\
A_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T} &:= \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)),
\end{aligned}
$$

for all $w \in [w(T)]$, where $\mathrm{D}_{\mathrm{KL}}(x_1\|x_2)$ for $x_1, x_2 \in [0,1]$ is the Kullback-Leibler divergence of the p.d.f. of two Bernoulli (formally defined in Appendix C.1).

Proof For $\boldsymbol{o} \in \{0, \dots, k\}^{\mathbb{N}_{\geqslant 1}}$, $t \in [T]$, let $i^*_{\boldsymbol{o},t} \in \arg\max_{i \in [k]} \mu_{\boldsymbol{o},i}(t)$. Then:

$$R_{\mathcal{E}_{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}}}}(\pi, T) = \sup_{\boldsymbol{o} \in \{0,\dots,k\}^{\mathbb{N}_{\geqslant 1}}} \mathop{\mathbb{E}}_{\boldsymbol{s} \sim \lambda} \left[\sum_{t=1}^{T} \left(\mu_{\boldsymbol{o},i^*_{\boldsymbol{o},t}}(t) - \mu_{\boldsymbol{o},I_t}(t) \right) \right]$$
$$\geqslant \mathop{\mathbb{E}}_{\substack{\boldsymbol{o} \sim \widetilde{\tau} \\ \boldsymbol{s} \sim \lambda}} \left[\sum_{t=1}^{T} \left(\mu_{\boldsymbol{o},i^*_{\boldsymbol{o},t}}(t) - \mu_{\boldsymbol{o},I_t}(t) \right) \right].$$

Under the assumption $\widetilde{\mu}_w(t) \geqslant \overline{\mu}_w(t)$ for all $w \in \mathbb{N}_{\geqslant 1}$, $t \in [D_w]$, we have:

$$\mu_{\boldsymbol{o},i^*_{\boldsymbol{o},t}}(t) - \mu_{\boldsymbol{o},I_t}(t) = \mathbf{1}[O_{w(t)} \neq 0, O_{w(t)} \neq I_t](\widetilde{\mu}_{w(t)}(t - s_{w(t)} + 1) - \overline{\mu}_{w(t)}(t - s_{w(t)} + 1)).$$

Then, observing that $O_w = 0$ has probability 0 under $\widetilde{\tau}$:

$$
\begin{aligned}
R_{\mathcal{E}_{\overline{\boldsymbol{\mu}},\tilde{\boldsymbol{\mu}}}}(\pi, T) &\geqslant \sum_{w=1}^{w(T)} \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)) \mathop{\mathbb{E}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[\mathbf{1}[O_w \neq I_t]] \\
&= \sum_{w=1}^{w(T)} \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)) \sum_{i\in[k]} \mathop{\mathbb{E}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[\mathbf{1}[I_t \neq i, O_w = i]] \\
&= \sum_{w=1}^{w(T)} \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)) \sum_{i\in[k]} \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[O_w = i] \frac{\mathbb{E}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[\mathbf{1}[I_t \neq i, O_w = i]]}{\mathbb{P}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[O_w = i]} \\
&= \sum_{w=1}^{w(T)} \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)) \frac{1}{k} \sum_{i\in[k]} \mathop{\mathbb{E}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[\mathbf{1}[I_t \neq i\,] | \, O_w = i] \\
&= \sum_{w=1}^{w(T)} \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)) \frac{1}{k} \sum_{i\in[k]} \left(1 - \mathop{\mathbb{E}}_{\substack{\boldsymbol{o}\sim\widetilde{\tau}\\ \boldsymbol{s}\sim\lambda}}[\mathbf{1}[I_t = i\,] | \, O_w = i]\right) \\
&\geqslant \sum_{w=1}^{w(T)} \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)) \frac{1}{k} \sum_{i\in[k]} \left(1 - \mathop{\mathbb{E}}_{\substack{\boldsymbol{o}\sim\overline{\tau}_w\\ \boldsymbol{s}\sim\lambda}}[\mathbf{1}[I_t = i]]\right. \\
&\quad \left. - \frac{1}{2}\sqrt{2\ln(2) \sum_{t'=s_w}^{\min\{e_w,T\}} \mathrm{D}_{\mathrm{KL}}(\overline{\mu}_w(t'-s_w+1) \| \widetilde{\mu}_w(t'-s_w+1)) \mathop{\mathbb{P}}_{\substack{\boldsymbol{o}\sim\overline{\tau}_w\\ \boldsymbol{s}\sim\lambda}}[I_{t'} = i]}\right) \\
&\geqslant \sum_{w=1}^{w(T)} \sum_{t=s_w}^{\min\{e_w,T\}} (\widetilde{\mu}_w(t-s_w+1) - \overline{\mu}_w(t-s_w+1)) \left(1 - \frac{1}{k}\right. \\
&\quad \left. - \frac{\sqrt{k}}{2k}\sqrt{2\ln(2) \sum_{t'=s_w}^{\min\{e_w,T\}} \mathrm{D}_{\mathrm{KL}}(\overline{\mu}_w(t'-s_w+1) \| \widetilde{\mu}_w(t'-s_w+1))}\right),
\end{aligned}
$$

(C.4)

where line (C.4) follows from Lemma 45 with f corresponding to the function from the observed rewards to the arm I_t pulled in round t, which is well defined for deterministic policies, and line (C.5) follows from Cauchy-Schwarz inequality applied to a vector of k ones and the vector of the terms under square root. The result follows from the definitions of $D_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}$ and $A_w^{\overline{\boldsymbol{\mu}},\widetilde{\boldsymbol{\mu}},T}$. ■

C.1.2. Specializing the Lower Bound for the Rising Setting

The goal of this section is to prove Theorem 18.

Theorem 18 (Lower Bound for the Rising Setting). *For any deterministic policy π and learning horizon $T \in \mathbb{N}_{\geqslant 1}$, $T \geqslant 2^{-3} k \min\{1, V_T\}^{-2}$, assuming Bernoulli-distributed rewards, it holds that:*

$$\sup_{\boldsymbol{\nu} \in \mathcal{E}_{\mathrm{r}}(T, V_T)} R_{\boldsymbol{\nu}}(\pi, T) \geqslant \frac{1}{80} T^{\frac{2}{3}} k^{\frac{1}{3}} \min\{1, V_T\}^{\frac{1}{3}}.$$

Proof First of all, we need to formally define the sequences of window widths, base, and modified trends. Let $D_{\mathrm{r},w} = D_{\mathrm{r}} := \left\lfloor T^{2/3} k^{1/3} \min\{1, V_T\}^{-2/3} \right\rfloor$and:

$$\overline{\mu}_{\mathrm{r},w}(t) := \begin{cases} \delta_{\mathrm{r}} + \varepsilon_{\mathrm{r}}(w-1) & \text{if } w \leqslant w(T) \\ \delta_{\mathrm{r}} + \varepsilon_{\mathrm{r}} w(T) & \text{if } w > w(T) \end{cases},$$

$$\widetilde{\mu}_{\mathrm{r},w}(t) := \begin{cases} \delta_{\mathrm{r}} + \varepsilon_{\mathrm{r}} w & \text{if } w \leqslant w(T) \\ \delta_{\mathrm{r}} + \varepsilon_{\mathrm{r}} w(T) & \text{if } w > w(T) \end{cases},$$

for all $w \in \mathbb{N}_{\geqslant 1}$ where $\delta_{\mathrm{r}} := \frac{2}{5}$ and $\varepsilon_{\mathrm{r}} := (1 - 2\delta_{\mathrm{r}}) \min\{1, V_T\} / w(T) > 0$. Observe that $\widetilde{\mu}_{\mathrm{r},w}(D_{\mathrm{r}}) \leqslant \overline{\mu}_{\mathrm{r},w+1}(1)$ for all $w \in \mathbb{N}_{\geqslant 1}$, hence, for any choice of $\boldsymbol{o} \in \{0, \dots, k\}^{\mathbb{N}_{\geqslant 1}}$, $\boldsymbol{\nu}_{\mathrm{r},\boldsymbol{o}}$ satisfies Assumption 4. Furthermore, for all $\boldsymbol{o} \in \{0, \dots, k\}^{\mathbb{N}_{\geqslant 1}}$, the expected rewards of the arms change at most between one window and the next, *i.e.*, $w(T) - 1$ times in the learning horizon, and the magnitude of the increment is at most $2\varepsilon_{\mathrm{r}}$, thus:

$$\Upsilon_{\boldsymbol{\nu}_{\mathrm{r},\boldsymbol{o}}}(1, T) \leqslant 2(w(T) - 1)\varepsilon_{\mathrm{r}} \leqslant V_T.$$

Hence $\mathcal{E}_{\overline{\boldsymbol{\mu}}_{\mathrm{r}}, \widetilde{\boldsymbol{\mu}}_{\mathrm{r}}} \subseteq \mathcal{E}_{\mathrm{r}}(T, V_T)$ indeed holds. Finally, it is easy to verify that $0 \leqslant \overline{\mu}_{\mathrm{r},w}(t) \leqslant \widetilde{\mu}_{\mathrm{r},w}(t) \leqslant 1$ for all $w \in \mathbb{N}_{\geqslant 1}$, $t \in [D_{\mathrm{r}}]$, so that the assumptions of Lemma 17 are satisfied. From Lemma 57, we have that:

$$D_w^{\overline{\boldsymbol{\mu}}_{\mathrm{r}}, \widetilde{\boldsymbol{\mu}}_{\mathrm{r}}, T} \leqslant \frac{\varepsilon_{\mathrm{r}}^2 D_{\mathrm{r}}}{2 \ln(2) \delta_{\mathrm{r}}^2}.$$

The choice of D_{r} implies $\varepsilon_{\mathrm{r}} \leqslant (1-2\delta_{\mathrm{r}})T^{-\frac{1}{3}}k^{\frac{1}{3}}\min\{1,V_T\}^{\frac{1}{3}}$. Then:

$$D_w^{\overline{\boldsymbol{\mu}}_{\mathrm{r}},\tilde{\boldsymbol{\mu}}_{\mathrm{r}},T} \leqslant \frac{k}{2\ln(2)}\left(\frac{1-2\delta_{\mathrm{r}}}{\delta_{\mathrm{r}}}\right)^2 \text{ for all } w \in [w(T)].$$

Thus, because of the choice of δ_{r} and observing that $k \geqslant 2$, we have

$$1 - \frac{1}{k} - \frac{\sqrt{2k}}{2k}\sqrt{\ln(2)D_w^{\overline{\boldsymbol{\mu}}_{\mathrm{r}},\tilde{\boldsymbol{\mu}}_{\mathrm{r}},T}} \geqslant \frac{1}{4}.$$

Since $A_w^{\overline{\boldsymbol{\mu}}_{\mathrm{r}},\tilde{\boldsymbol{\mu}}_{\mathrm{r}},T} = \varepsilon_{\mathrm{r}}(\min\{e_w,T\} - s_w + 1)$, by plugging the previous results in Lemma 17, assuming that $T \geqslant 2^{-3}k\min\{1,V_T\}^{-2}$:

$$\sup_{\boldsymbol{\nu}\in\mathcal{E}_{\mathrm{r}}(T,V_T)} R_{\boldsymbol{\nu}}(\pi,T) \geqslant \frac{1}{4}\varepsilon_{\mathrm{r}}T \geqslant \frac{1}{80}T^{\frac{2}{3}}k^{\frac{1}{3}}\min\{1,V_T\}^{\frac{1}{3}},$$

where the last step follows from the definition of ε_{r} and the fact that $\lfloor x \rfloor \geqslant x/2$ and $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$. ■

C.1.3. Specializing the Lower Bound for the Rising Concave Setting

The goal of this section is to prove Theorem 19.

Theorem 19 (Lower Bound for the Rising Concave Setting)**.** *For any deterministic policy π and learning horizon $T \in \mathbb{N}_{\geqslant 1}$, $T \geqslant 2^{10}k\min\{1,V_T\}^{-2}$, with Bernoulli-distributed rewards, it holds:*

$$\sup_{\boldsymbol{\nu}\in\mathcal{E}_{\mathrm{c}}(T,V_T)} R_{\boldsymbol{\nu}}(\pi,T) \geqslant 2^{-15}T^{\frac{3}{5}}k^{\frac{2}{5}}\min\{1,V_T\}^{\frac{1}{5}}.$$

Proof First of all, we need to formally define the sequences of window widths, base, and modified trends. Let $N_{\mathrm{c}} \coloneqq \left\lceil T^{1/5}k^{-1/5}\min\{1,V_T\}^{2/5}\right\rceil$, $D_{\mathrm{c},w} = D_{\mathrm{c}} \coloneqq \lceil T/N_{\mathrm{c}} \rceil$ for all $w \in \mathbb{N}_{\geqslant 1}$. Observe that D_{c} is defined in such a way that $w(T) \leqslant N_{\mathrm{c}}$. Let $\delta_{\mathrm{c}} \coloneqq 1/3$, $m_0 \coloneqq (1-2\delta_{\mathrm{c}})\min\{1,V_T\}/(2T) \in (0,1)$, $m_w \coloneqq (2N_{\mathrm{c}} - w)m_0/(2N_{\mathrm{c}})$ for $w \in [2N_{\mathrm{c}}]$. $(m_w)_{w=0}^{2N_{\mathrm{c}}}$ are the slopes of the segments which constitute the trends. Observe that $m_0 > m_1 > \cdots > m_{2N_{\mathrm{c}}-1} >$

$m_{2N_c} = 0$. We are ready to define the trends:

$$\overline{\mu}_{c,w}(t) := \begin{cases} \delta_c + D_c \sum_{l=1}^{w-1} m_{2l-1} + t m_{2w-1} & \text{if } w \leqslant w(T) \\ \delta_c + D_c \sum_{l=1}^{w(T)} m_{2l-1} & \text{if } w > w(T) \end{cases},$$

$$\widetilde{\mu}_{c,w}(t) := \begin{cases} \delta_c + D_c \sum_{l=1}^{w-1} m_{2l-1} + t m_{2w-2} + \left(t - \frac{D_c}{2}\right)^+ (m_{2w} - m_{2w-2}) & \text{if } w \leqslant w(T) \\ \delta_c + D_c \sum_{l=1}^{w(T)} m_{2l-1} & \text{if } w > w(T) \end{cases},$$

for all $w \in \mathbb{N}_{\geqslant 1}$. In what follows, with a slight abuse of notation, we will regard $\overline{\mu}_{c,w}$ and $\widetilde{\mu}_{c,w}$ as defined on $[0, D_c]$. Observe that, as we informally stated before, $\overline{\mu}_{c,w}(0) = \widetilde{\mu}_{c,w}(0)$, and $\overline{\mu}_{c,w}(D_c) = \widetilde{\mu}_{c,w}(D_c) = \overline{\mu}_{c,w+1}(0)$ for all $w \in \mathbb{N}_{\geqslant 1}$. Furthermore, it is easy to check that the slope of the second segment of the modified trend in a window is equal to the slope of the first segment of the modified trend in the next window. Thus, because of what we remarked when we informally introduced the construction, for any choice of $\boldsymbol{o} \in \{0, \dots, k\}^{\mathbb{N}_{\geqslant 1}}$, $\boldsymbol{\nu}_{c,\boldsymbol{o}}$ satisfies Assumptions 4 and 5. Furthermore, in each window with index $w \in [w(T)]$, the maximum increment of the expected reward of an arm, corresponds to the slope of the first half of the modified trend m_{2w-2}. Thus:

$$\Upsilon_{\boldsymbol{\nu}_{c,\boldsymbol{o}}}(1, T) \leqslant D_c \sum_{w=1}^{w(T)} m_{2w-2} \leqslant D_c \sum_{w=1}^{N_c} m_{2w-2} \leqslant V_T,$$

because of how we defined the quantities involved. Hence $\mathcal{E}_{\overline{\boldsymbol{\mu}}_c, \widetilde{\boldsymbol{\mu}}_c} \subseteq \mathcal{E}_c(T, V_T)$ indeed holds. Finally, it is easy to verify that $0 \leqslant \overline{\mu}_{c,w}(t) \leqslant \widetilde{\mu}_{c,w}(t) \leqslant 1$ for all $w \in \mathbb{N}_{\geqslant 1}$, $t \in [D_c]$, so that the assumptions of Lemma 17 are satisfied. The maximum distance between the two trends in a window is attained for $t = \frac{D_c}{2}$ and has value:

$$\varepsilon_c := \widetilde{\mu}_{c,w}\left(\frac{D_c}{2}\right) - \overline{\mu}_{c,w}\left(\frac{D_c}{2}\right) = \frac{D_c m_0}{4N_c}.$$

Because of how D_c and m_0 are defined, remembering that $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$, we have:

$$\varepsilon_c \leqslant \frac{(1 - 2\delta_c)\min\{1, V_T\}}{4N_c^2},$$

hence, in virtue of Lemma 57, we have:

$$D_w^{\overline{\boldsymbol{\mu}}_{\mathrm{c}},\tilde{\boldsymbol{\mu}}_{\mathrm{c}},T} \leqslant \frac{(1-2\delta_{\mathrm{c}})^2 \min\{1,V_T\}^2}{32N_{\mathrm{c}}^4 \ln(2)\delta_{\mathrm{c}}^2} D_{\mathrm{c}} \leqslant \frac{(1-2\delta_{\mathrm{c}})^2 \min\{1,V_T\}^2}{16\delta_{\mathrm{c}}^2 N_{\mathrm{c}}^5 \ln(2)} T \quad \text{for all} \quad w \in [w(T)].$$

Thus, under our choices of N_{c} and δ_{c}, remembering that $k \geqslant 2$, we get:

$$1 - \frac{1}{k} - \frac{\sqrt{2k}}{2k}\sqrt{\ln(2) D_w^{\overline{\boldsymbol{\mu}}_{\mathrm{c}},\tilde{\boldsymbol{\mu}}_{\mathrm{c}},T}} \geqslant \frac{1}{4}.$$

Now, let's lower bound the expression of $A_w^{\overline{\boldsymbol{\mu}}_{\mathrm{c}},\tilde{\boldsymbol{\mu}}_{\mathrm{c}},T}$ for $w \in [w(T)-1]$:

$$\begin{aligned}
A_w^{\overline{\boldsymbol{\mu}}_{\mathrm{c}},\tilde{\boldsymbol{\mu}}_{\mathrm{c}},T} &= \sum_{t=1}^{\left\lfloor \frac{D_{\mathrm{c}}}{2} \right\rfloor} (m_{2w-2} - m_{2w-1})t + \sum_{t=\left\lfloor \frac{D_{\mathrm{c}}}{2} \right\rfloor + 1}^{D_{\mathrm{c}}} \left[(m_{2w} - m_{2w-1})t - \frac{D_{\mathrm{c}}}{2}(m_{2w} - m_{2w-2}) \right] \\
&= \frac{m_0}{2N_{\mathrm{c}}} \left(\frac{\left\lfloor \frac{D_{\mathrm{c}}}{2} \right\rfloor \left(\left\lfloor \frac{D_{\mathrm{c}}}{2} \right\rfloor + 1 \right)}{2} + \frac{\left(D_{\mathrm{c}} - \left\lfloor \frac{D_{\mathrm{c}}}{2} \right\rfloor - 1 \right) \left(D_{\mathrm{c}} - \left\lfloor \frac{D_{\mathrm{c}}}{2} \right\rfloor \right)}{2} \right) \\
&\geqslant \frac{m_0}{4N_{\mathrm{c}}} \left(\frac{5}{16} D_{\mathrm{c}}^2 - \frac{D_{\mathrm{c}}}{4} \right) \geqslant \frac{m_0}{4N_{\mathrm{c}}} \left(\frac{5}{16} D_{\mathrm{c}}^2 - \frac{4}{16} D_{\mathrm{c}}^2 \right) = \frac{m_0 D_{\mathrm{c}}^2}{64 N_{\mathrm{c}}} \\
&\geqslant \frac{m_0 T^2}{64 N_{\mathrm{c}}^3} = \frac{(1-2\delta_{\mathrm{c}}) \min\{1,V_T\} T}{128 N_{\mathrm{c}}^3},
\end{aligned} \tag{C.6}$$

where line (C.6) follows from the fact that $\lfloor x \rfloor \geqslant x/2$ for $x \geqslant 1$ and that $x \leqslant x^2$ for $x \geqslant 1$. Finally, $T \geqslant 2^{10} k \min\{1,V_T\}^{-2}$ guarantees $w(T) - 1 \geqslant N_{\mathrm{c}}/4$, and thus, by Lemma 17 in conjunction with the results we just proved, we have:

$$\begin{aligned}
\sup_{\boldsymbol{\nu} \in \mathcal{E}_{\mathrm{c}}(T,V_T)} R_{\boldsymbol{\nu}}(\pi,T) &\geqslant \frac{1}{4} \sum_{w=1}^{w(T)} A_w^{\overline{\boldsymbol{\mu}}_{\mathrm{c}},\tilde{\boldsymbol{\mu}}_{\mathrm{c}},T} \geqslant \frac{1}{4} \left(w(T) - 1 \right) \frac{(1-2\delta_{\mathrm{c}}) \min\{1,V_T\} T}{128 N_{\mathrm{c}}^3} \\
&\geqslant \frac{(1-2\delta_{\mathrm{c}}) \min\{1,V_T\} T}{2^{11} N_{\mathrm{c}}^2} \geqslant 2^{-15} T^{\frac{3}{5}} k^{\frac{2}{5}} \min\{1,V_T\}^{\frac{1}{5}},
\end{aligned} \tag{C.7}$$

where line (C.7) follows from our choices of N_{c} and δ_{c} and from the fact that $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$.
■

C.2. Upper Bound for the Rising Concave Setting

In this appendix, we provide the proofs of the results presented in Section 3.3.5 in the main paper.

C.2.1. Additional notation

We begin by introducing the additional notation required for the analysis. Let

$$\hat{S}_{i,w,d} \coloneqq \sum_{t=s_w^{(\alpha)}}^{s_w^{(\alpha)}+d-1} \mathbf{1}[I_t = i] R_t, \qquad \widetilde{S}_{i,w,d} \coloneqq \sum_{t=s_w^{(\alpha)}}^{s_w^{(\alpha)}+d-1} \mathbf{1}[I_t = i] \mu_i(t)$$

be respectively the *cumulative reward* and *cumulative expected reward* by $\texttt{RC-BE}(\alpha)$ for arm $i \in [k]$ in the first $d \in \{0, \dots, D_w^{(\alpha)}\}$ rounds of window $w \in \mathbb{N}_{\geqslant 1}$. Let N_w be the number of round-robin cycles of window $w \in \mathbb{N}_{\geqslant 1}$, where we also count the degenerate cycles in which we pull the only remaining alive arm $\hat{i}^*$. Let $t_{w,l}$ be the round in which the l-th round-robin cycle (with $l \in [N_w]$) is started during window $w \in \mathbb{N}_{\geqslant 1}$. Analogously, let $N_{i,w}$ be the number of times arm $i \in [k]$ is pulled in the w-th window (with $w \in \mathbb{N}_{\geqslant 1}$) and $t_{i,w,l}$ the round in which arm i is pulled for the l-th time (with $l \in [N_{i,w}]$) during window w. For simplicity in the notation, we define $d_{w,l} = t_{w,l} - s_w^{(\alpha)} + 1$ and $d_{i,w,l} = t_{i,w,l} - s_w^{(\alpha)} + 1$. Finally we define the good events

$$\mathcal{G}_{i,w,d,\delta} \coloneqq \left\{ \left| \hat{S}_{i,w,d} - \widetilde{S}_{i,w,d} \right| \leqslant \sigma \sqrt{2 D_w^{(\alpha)} \left(\ln\left(2k D_w^{(\alpha)}\right) + \ln\left(\frac{1}{\delta}\right) \right)} \right\}$$

for $i \in [k]$, $w \in \mathbb{N}_{\geqslant 1}$, $d \in [D_w^{(\alpha)}]$, $\delta \in (0, 1]$, and

$$\mathcal{G}_{w,\delta} = \bigcap_{\substack{i \in [k] \\ d \in [D_w^{(\alpha)}]}} \mathcal{G}_{i,w,d,\delta}$$

for $i \in [k]$, $\delta \in (0, 1]$.

C.2.2. Concentration

We start the analysis with a concentration result for $\hat{S}_{i,w,d}$.

Lemma 46 (Concentration)**.** *For every $w \in \mathbb{N}_{\geqslant 1}$, $\delta \in (0, 1]$, we have that:*

$$\mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\overline{\mathcal{G}_{w,\delta}}\right] \leqslant \delta.$$

Proof For $i \in [k]$, $d \in \{0, \dots, D_w^{(\alpha)}\}$, $\lambda \in \mathbb{R}$, let:

$$M_{i,w,d}(\lambda) \coloneqq \exp\left(\lambda\left(\hat{S}_{i,w,d} - \widetilde{S}_{i,w,d}\right)\right),$$

$$\mathcal{F}_{w,d} \coloneqq \sigma\left(X_{1,1}, \dots, X_{k,1}, X_{1,s_w^{(\alpha)}+d-1}, \dots, X_{k,s_w^{(\alpha)}+d-1}\right).$$

Let $t' \coloneqq s_w^{(\alpha)} + d - 1$ to ease the notation. Observe that $I_{t'}$ is $\mathcal{F}_{w,d-1}$-measurable and that $X_{i,t'}$ is independent from $\mathcal{F}_{w,d-1}$. Furthermore, we can rewrite $\hat{S}_{i,w,d}$ as

$$\hat{S}_{i,w,d} = \sum_{t=s_w^{(\alpha)}}^{t'} \mathbf{1}[I_t = i] X_{i,t}.$$

Then:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[M_{i,w,d}(\lambda) \mid \mathcal{F}_{w,d-1}\right] &= M_{i,w,d-1}(\lambda) \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\mathbf{1}[I_{t'} = i] \exp\left(\lambda\left(X_{i,t'} - \mu_i(t')\right)\right)\right. \\
&\quad \left. + 1 - \mathbf{1}[I_{t'} = i] \mid \mathcal{F}_{w,d-1}\right] \\
&\leqslant M_{i,w,d-1}(\lambda) \exp\left(\mathbf{1}[I_{t'} = i]\frac{\lambda^2\sigma^2}{2}\right) \leqslant M_{i,w,d-1}(\lambda) \exp\left(\frac{\lambda^2\sigma^2}{2}\right),
\end{aligned}$$

where in the last line we use the properties of conditional expectation (Klenke, 2020) and the sub-gaussianity of $X_{i,t'}$. Thus, by induction:

$$\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[M_{i,w,d}(\lambda)\right] \leqslant \exp\left(d\frac{\lambda^2\sigma^2}{2}\right) \leqslant \exp\left(D_w^{(\alpha)}\frac{\lambda^2\sigma^2}{2}\right).$$

Then, thanks to Markov inequality, for every $\varepsilon \in \mathbb{R}$:

$$\begin{aligned}
\mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\hat{S}_{i,w,d} - \widetilde{S}_{i,w,d} > \varepsilon\right] &= \mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[M_{i,w,d}(\lambda) > \exp(\lambda\varepsilon)\right] \\
&\leqslant \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[M_{i,w,d}(\lambda)\right]\exp(-\lambda\varepsilon) \\
&\leqslant \exp\left(\lambda^2\frac{D_w^{(\alpha)}\sigma^2}{2} - \lambda\varepsilon\right).
\end{aligned}$$

By choosing $\varepsilon = \sigma\sqrt{2D_w^{(\alpha)}\left(\ln\left(2kD_w^{(\alpha)}\right) + \ln\left(\frac{1}{\delta}\right)\right)}$, $\lambda = \frac{\varepsilon}{D_w^{(\alpha)}\sigma^2}$, we get:

$$\mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\hat{S}_{i,w,d} - \widetilde{S}_{i,w,d} > \varepsilon\right] \leqslant \frac{\delta}{2kD_w^{(\alpha)}}.$$

An analogous bound holds for

$$\mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\widetilde{S}_{i,w,d} - \hat{S}_{i,w,d} > \varepsilon\right].$$

Then, thanks to a union bound,

$$\mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\overline{\mathcal{G}_{i,w,d,\delta}}\right] \leqslant \frac{\delta}{kD_w^{(\alpha)}}.$$

Finally:

$$\mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\overline{\mathcal{G}_{w,\delta}}\right] \leqslant \sum_{i\in[k]}\sum_{d\in[D_w^{(\alpha)}]}\mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\overline{\mathcal{G}_{i,w,d,\delta}}\right] \leqslant \delta.$$

■

C.2.3. Proof of Lemma 20

The goal of this section is to prove Lemma 20. To this end, we need several intermediate results. We start by proving that $\mathcal{I}_w^{\times}$ is indeed well-defined.

Lemma 47. *Let $i_w^*, j_w^* \in \mathcal{I}_w^*$, then $i_w^* \,{}_w\!\times j_w^*$.*

Proof If $\mu_{i_w^*}(t') = \mu_{j_w^*}(t')$ for some $t' \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$, then it must be $i_w^* \,{}_w\!\times j_w^*$. Thus, assume $\mu_{i_w^*}(t') < \mu_{j_w^*}(t')$ for some $t' \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$. If i_w^* and j_w^* do not cross, then

$$\mu_{i_w^*}(t) < \mu_{j_w^*}(t) \text{ for all } t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$$

which is a contradiction with the fact that $i_w^* \in \mathcal{I}_w^*$. ■

We now prove a very useful property of ${}_w\!\times^+$.

Lemma 48. *Let $i, j, k \in [k]$. If $i \,{}_w\!\times^+ j$ and there exists $t' \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$ such that $\mu_i(t') \leqslant \mu_k(t') \leqslant \mu_j(t')$, then $k \in [i]_{{}_w\times^+}$.*

Proof If $\mu_k(t') = \mu_i(t')$ or $\mu_k(t') = \mu_j(t')$ then the statement is trivial. Consider $\mu_i(t') <$

$\mu_k(t') < \mu_j(t')$. We proceed by contradiction. Assume that it is not true that $k \,{}_w\!\times^+ i$. Let $\mathcal{I}_1 = \left\{l \in [i]_{w\times^+} \text{ s.t. } \mu_l(t') < \mu_k(t')\right\}$ and $\mathcal{I}_2 = \left\{l \in [i]_{w\times^+} \text{ s.t. } \mu_l(t') > \mu_k(t')\right\}$. Since $\mathcal{I}_1 \cup \mathcal{I}_2 \subseteq [i]_{w\times^+}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \{\}$, $\mathcal{I}_1, \mathcal{I}_2 \neq \{\}$ there must be $i_1 \in \mathcal{I}_1$, $i_2 \in \mathcal{I}_2$ such that $i_1 \,{}_w\!\times i_2$. But, since it is not true that $k \,{}_w\!\times^+ i$, it cannot be $k \,{}_w\!\times i_1$ nor $k \,{}_w\!\times i_2$. Thus it must be

$$\mu_{i_1}(t) < \mu_k(t) < \mu_{i_2}(t) \text{ for all } t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}.$$

But this is absurd since $i_1 \,{}_w\!\times i_2$, concluding the proof. ■

This leads to the following corollary.

Corollary 49. *Let $i \in \mathcal{I}_w^\times$, $j \notin \mathcal{I}_w^\times$, then:*

$$\mu_j(t) < \mu_i(t) \textit{ for all } t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}.$$

Proof By contrapositive, if $\mu_j(t') \geqslant \mu_i(t')$ for some $t' \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$, then there exists $k \in \arg\max_{l\in[k]} \mu_l(t')$ such that $\mu_i(t') \leqslant \mu_j(t') \leqslant \mu_k(t')$ and thus $j \in \mathcal{I}_w^\times$ by Lemma 48. ■

We are ready to prove Lemma 20.

Lemma 20. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $w \in \mathbb{N}_{\geqslant 1}$ we have that:*

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \{w\}) \leqslant \underbrace{3kB_w^{(\alpha)}}_{\text{Exploration}} + \underbrace{D_w^{(\alpha)} d_w^*}_{\text{Commitment}}.$$

Proof We start by proving that, under event $\mathcal{G}_{w,(2kD_w^{(\alpha)})^{-1}}$, at least one arm in $\mathcal{I}_w^\times$ is always alive in each round-robin cycle. We need to consider all the eliminations which happen at the end of a round-robin cycle, except for the last, in which eliminations are irrelevant (remember that the window ends at the end of the last round-robin cycle and the algorithm is restarted). To this end, let $n \in [N_w - 1]$. For an arm $i \in [k]$, to eliminate an arm $j \in [k]$ at the end of the n-th round-robin cycle, it must be:

$$\hat{S}_{i,w,d_{w,n+1}-1} > \hat{S}_{j,w,d_{w,n+1}-1} + B_w^{(\alpha)}$$

which, under event $\mathcal{G}_{w,(2kD_w^{(\alpha)})^{-1}}$, implies

$$\widetilde{S}_{i,w,d_{w,n+1}-1} + 4\sigma\sqrt{D_w^{(\alpha)} \ln\left(2kD_w^{(\alpha)}\right)} > \widetilde{S}_{j,w,d_{w,n+1}-1} + B_w^{(\alpha)}$$

if and only if

$$\sum_{l=1}^{n}[\mu_i(t_{w,l}) + \mu_i(t_{i,w,l}) - \mu_i(t_{w,l})] + 4\sigma\sqrt{D_w^{(\alpha)}\ln\left(2kD_w^{(\alpha)}\right)}$$
$$> \sum_{l=1}^{n}[\mu_j(t_{w,l}) + \mu_j(t_{j,w,l}) - \mu_j(t_{w,l})] + B_w^{(\alpha)}$$

which implies, being the instance rising:

$$\sum_{l=1}^{n}\mu_i(t_{w,l}) + 1 + 4\sigma\sqrt{D_w^{(\alpha)}\ln\left(2kD_w^{(\alpha)}\right)} > \sum_{l=1}^{n}\mu_j(t_{w,l}) + B_w^{(\alpha)}$$

and thus, because of the choice of $B_w^{(\alpha)}$, it must be:

$$\sum_{l=1}^{n}\mu_i(t_{w,l}) > \sum_{l=1}^{n}\mu_j(t_{w,l}).$$

Thus, in virtue of Corollary 49, it cannot be $i \notin \mathcal{I}_w^{\times}$, $j \in \mathcal{I}_w^{\times}$. But, to eliminate all alive arms in $\mathcal{I}_w^{\times}$, we would need at least one cycle in which an elimination of the kind above happens. Hence there will always be at least an arm in $\mathcal{I}_w^{\times}$ alive. Let $i_{w,n}^{\times}$ be such arm during the n-th round-robin cycle. Let's bound the regret of a generic arm $j \in [k]$ during the w-th window, under event

$\mathcal{G}_{w,(2kD_w^{(\alpha)})^{-1}}$.

$$\begin{aligned}
\sum_{l=1}^{N_{j,w}} \left[\mu_{i^*_{t_{j,w,l}}}(t_{j,w,l}) - \mu_j(t_{j,w,l})\right] &\leqslant \sum_{l=1}^{N_{j,w}-1} \left[\mu_{i^*_{t_{j,w,l}}}(t_{j,w,l}) - \mu_j(t_{j,w,l})\right] + 1 \\
&\leqslant \sum_{l=1}^{N_{j,w}-1} \left[\mu_{i^\times_{w,N_{j,w}}}(t_{j,w,l}) - \mu_j(t_{j,w,l})\right] + N_{j,w}d^*_w + 1 \\
&= \sum_{l=1}^{N_{j,w}-1} \left[\mu_{i^\times_{w,N_{j,w}}}(t_{i^\times_{w,N_{j,w}},w,l}) - \mu_j(t_{j,w,l})\right] \\
&+ \sum_{l=1}^{N_{j,w}-1} \left[\mu_{i^\times_{w,N_{j,w}}}(t_{j,w,l}) - \mu_{i^\times_{w,N_{j,w}}}(t_{i^\times_{w,N_{j,w}},w,l})\right] \\
&+ N_{j,w}d^*_w + 1 \\
&\leqslant \widetilde{S}_{i^\times_{w,N_{j,w}},w,d_{w,N_{j,w}}-1} - \widetilde{S}_{j,w,d_{w,N_{j,w}}-1} + 1 + N_{j,w}d^*_w + 1 \\
&\leqslant 2 + 4\sigma\sqrt{D_w^{(\alpha)} \ln\left(2kD_w^{(\alpha)}\right)} + \hat{S}_{i^\times_{w,N_{j,w}},w,d_{w,N_{j,w}}-1} \\
&- \hat{S}_{j,w,d_{w,N_{j,w}}-1} + N_{j,w}d^*_w \\
&= B_w^{(\alpha)} + \hat{S}_{i^\times_{w,N_{j,w}},w,d_{w,N_{j,w}}-1} - \hat{S}_{j,w,d_{w,N_{j,w}}-1} + N_{j,w}d^*_w \\
&\leqslant 2B_w^{(\alpha)} + N_{j,w}d^*_w
\end{aligned}$$

where the last line follows from the fact that we have not eliminated arm j at the end of the $(N_{j,w}-1)$-th round robin cycle. Thus, the regret during the w-th window, under event $\mathcal{G}_{w,(2kD_w^{(\alpha)})^{-1}}$, is upper bounded as:

$$\sum_{t=s_w^{(\alpha)}}^{e_w^{(\alpha)}} \left[\mu_{i_t^*}(t) - \mu_{I_t}(t)\right] = \sum_{j\in[k]} \sum_{l=1}^{N_{j,w}} \left[\mu_{i^*_{t_{j,w,l}}}(t_{j,w,l}) - \mu_j(t_{j,w,l})\right] \leqslant 2kB_w^{(\alpha)} + D_w^{(\alpha)}d^*_w.$$

Finally, in virtue of Lemma 46:

$$\begin{aligned}
R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \{w\}) &\leqslant 2kB_w^{(\alpha)} + D_w^{(\alpha)}d^*_w + D_w^{(\alpha)} \mathop{\mathbb{P}}_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[\overline{\mathcal{G}_{w,(2kD_w^{(\alpha)})^{-1}}}\right] \\
&\leqslant 2kB_w^{(\alpha)} + D_w^{(\alpha)}d^*_w + \frac{1}{2k} \leqslant 3kB_w^{(\alpha)} + D_w^{(\alpha)}d^*_w.
\end{aligned}$$

■

C.2.4. Proof of Lemma 21

The goal of this section is to prove Lemma 21. To this end, we need several intermediate results. We start with a lower bound to $e_w^{(\alpha)}$.

Lemma 50. *For any $\alpha \geqslant 1$, $w \in \mathbb{N}_{\geqslant 1}$ it holds that*

$$e_w^{(\alpha)} \geqslant \frac{w^{1+\alpha}}{2(1+\alpha)}.$$

Proof If $w = 1$, we trivially have

$$e_1^{(\alpha)} = 1 > \frac{1}{2(1+\alpha)}.$$

Now, suppose $w \geqslant 2$, then

$$e_w^{(\alpha)} = \sum_{l=1}^{w} D_l^{(\alpha)} \geqslant \sum_{l=1}^{w} l^\alpha \geqslant \int_1^w x^\alpha dx = \left(\frac{w^{1+\alpha}}{1+\alpha} - \frac{1}{1+\alpha}\right) \geqslant \frac{w^{1+\alpha}}{2(1+\alpha)}.$$

■

Now we introduce the results through which we exploit the concavity of the instance.

Lemma 51. *For any restless rising concave MAB $\boldsymbol{\nu}$, $t_1, t_2 \in \mathbb{N}_{\geqslant 1}$, $t_2 \geqslant t_1 \geqslant 2$, we have:*

$$\Upsilon_{\boldsymbol{\nu}}(t_1, t_2) \leqslant \frac{t_2 - t_1}{t_2 - t_1 + 1} \Upsilon_{\boldsymbol{\nu}}(t_1 - 1, t_2).$$

Proof

$$\begin{aligned}
\Upsilon_{\boldsymbol{\nu}}(t_1, t_2) &= \sum_{l=t_1}^{t_2-1} \max_{i \in [k]} \gamma_i(l) \\
&\leqslant \sum_{l=t_1}^{t_2-1} \max_{i \in [k]} \gamma_i(l) + \frac{t_2 - t_1}{t_2 - t_1 + 1} \left(\max_{i \in [k]} \gamma_i(t_1 - 1) - \frac{1}{t_2 - t_1} \sum_{l=t_1}^{t_2-1} \max_{i \in [k]} \gamma_i(l) \right) \\
&= \frac{t_2 - t_1}{t_2 - t_1 + 1} \sum_{l=t_1-1}^{t_2-1} \max_{i \in [k]} \gamma_i(l) = \frac{t_2 - t_1}{t_2 - t_1 + 1} \Upsilon_{\boldsymbol{\nu}}(t_1 - 1, t_2).
\end{aligned}$$

■

Before proving Lemma 21, we need an intermediate upper bound to $d_w(i)$.

Lemma 52. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $w \in \mathbb{N}_{\geqslant 1}$, $i \in [k]$, we have that:*

$$d_w(i) \leqslant (|[i]_{w\times^+}| - 1) \max_{\substack{j,k\in[i]_{w\times^+} \text{ s.t.} j\ {}_w\times k \\ t\in\{s_w^{(\alpha)},\dots,e_w^{(\alpha)}\}}} |\mu_j(t) - \mu_k(t)|.$$

Proof If $j\ {}_w\times^+ k$, there must exist distinct $i_1, \dots, i_n$ different from j and k ($n \in \{0, \dots, |[i]_{w\times^+}| - 2\}$) such that $j\ {}_w\times i_1$, $i_1\ {}_w\times i_2$, $\dots$, $i_{n-1}\ {}_w\times i_n$, $i_n\ {}_w\times k$. Then, for $t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$, we have:

$$\begin{aligned}
|\mu_j(t) - \mu_k(t)| &\leqslant |\mu_j(t) - \mu_{i_1}(t)| + |\mu_{i_1}(t) - \mu_{i_2}(t)| + \dots |\mu_{i_n}(t) - \mu_k(t)| \\
&\leqslant (n+1) \max_{\substack{j',k'\in[i]_{w\times^+} \text{ s.t.} j'\ {}_w\times k' \\ t'\in\{s_w^{(\alpha)},\dots,e_w^{(\alpha)}\}}} |\mu_{j'}(t') - \mu_{k'}(t')| \\
&\leqslant \left(|[i]_{w\times^+}| - 1\right) \max_{\substack{j',k'\in[i]_{w\times^+} \text{ s.t.} j'\ {}_w\times k' \\ t'\in\{s_w^{(\alpha)},\dots,e_w^{(\alpha)}\}}} |\mu_{j'}(t') - \mu_{k'}(t')|.
\end{aligned}$$

■

We are ready to prove Lemma 21.

Lemma 21. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $w \in \mathbb{N}_{\geqslant 1}$, $i \in [k]$, we have that:*

$$d_w(i) \leqslant 8(1+\alpha)\left(\left|[i]_{w\times^+}\right| - 1\right) w^{-1}\Upsilon_{\boldsymbol{\nu}}(1, e_w^{(\alpha)}) \leqslant 16\alpha k w^{-1}\Upsilon_{\boldsymbol{\nu}}(1, e_w^{(\alpha)}).$$

Proof Let $j \uparrow_{t'} k$ for some $j, k \in [i]_{w\times^+}$, $t' \in \{s_w^{(\alpha)} + 1, \dots, e_w^{(\alpha)}\}$. Let $t \geqslant t'$, $t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$, then

$$\mu_j(t) - \mu_k(t) \leqslant \mu_j(t) - \mu_k(t'-1) \leqslant \mu_j(t) - \mu_j(t'-1) \leqslant \Upsilon_{\boldsymbol{\nu}}\left(s_w^{(\alpha)}, e_w^{(\alpha)}\right),$$

$$\mu_k(t) - \mu_j(t) \leqslant \mu_k(t) - \mu_j(t') \leqslant \mu_k(t) - \mu_k(t') \leqslant \Upsilon_{\boldsymbol{\nu}}\left(s_w^{(\alpha)}, e_w^{(\alpha)}\right).$$

Analogously, if $t < t'$, we have

$$\mu_j(t) - \mu_k(t) \leqslant \mu_j(t'-1) - \mu_k(t) \leqslant \mu_k(t'-1) - \mu_k(t) \leqslant \Upsilon_{\boldsymbol{\nu}}\left(s_w^{(\alpha)}, e_w^{(\alpha)}\right),$$

$$\mu_k(t) - \mu_j(t) \leqslant \mu_k(t') - \mu_j(t) \leqslant \mu_j(t') - \mu_j(t) \leqslant \Upsilon_{\boldsymbol{\nu}}\left(s_w^{(\alpha)}, e_w^{(\alpha)}\right).$$

We conclude that, if $j\ {}_w\times k$, then $|\mu_j(t) - \mu_k(t)| \leqslant \Upsilon_{\boldsymbol{\nu}}\left(s_w^{(\alpha)}, e_w^{(\alpha)}\right)$ for all $t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$.

Thus, in virtue of Lemma 52, if $j \,{}_w\times^+ k$, then:

$$|\mu_j(t) - \mu_k(t)| \leqslant \left(|[i]_{w\times^+}| - 1\right) \Upsilon_{\boldsymbol{\nu}}\left(s_w^{(\alpha)}, e_w^{(\alpha)}\right).$$

For $w \geqslant 2$, by applying iteratively Lemma 51, we have

$$\begin{aligned}\Upsilon_{\boldsymbol{\nu}}\left(s_w^{(\alpha)}, e_w^{(\alpha)}\right) &\leqslant \frac{e_w^{(\alpha)} - s_w^{(\alpha)}}{e_w^{(\alpha)} - 1}\Upsilon_{\boldsymbol{\nu}}\left(1, e_w^{(\alpha)}\right) \leqslant 2\frac{D_w^{(\alpha)}}{e_w^{(\alpha)}}\Upsilon_{\boldsymbol{\nu}}\left(1, e_w^{(\alpha)}\right) \\ &\leqslant 8(1+\alpha)\frac{w^\alpha}{w^{1+\alpha}}\Upsilon_{\boldsymbol{\nu}}\left(1, e_w^{(\alpha)}\right) = 8(1+\alpha)w^{-1}\Upsilon_{\boldsymbol{\nu}}\left(1, e_w^{(\alpha)}\right)\end{aligned}$$

where in the last line we used Lemma 50, the fact that $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$, and the definition of $D_w^{(\alpha)}$. The same upper bound holds trivially for $w = 1$ since $s_1^{(\alpha)} = e_1^{(\alpha)} = 1$. ■

C.2.5. Proof of Lemma 22

The goal of this section is to prove Lemma 22. To get the result, we start by providing an upper bound to the number of times an arm i overtakes arm j and the expected rewards diverge by a quantity greater than $G > 0$. To this end, we need to prove two auxiliary results.

Lemma 53. *Let $t^\uparrow, \hat{t}, t^\downarrow \in \mathbb{N}_{\geqslant 1}$, $t^\downarrow > \hat{t} \geqslant t^\uparrow$, $G \in (0, 1]$, $i, j \in [k]$ such that*

$$i \uparrow_{t^\uparrow} j, \mu_i(\hat{t}) \geqslant \mu_j(\hat{t}) + G, j \uparrow_{t^\downarrow} i.$$

Then:

$$\gamma_i(t^\uparrow - 1) > \gamma_j(\hat{t}) \geqslant \gamma_i(t^\downarrow), \tag{C.8}$$

$$\hat{t} - (t^\uparrow - 1) \geqslant G\frac{1}{\gamma_i(t^\uparrow - 1) - \gamma_j(\hat{t})}, \tag{C.9}$$

$$\mu_i(\hat{t}) - \mu_i(t^\uparrow - 1) \geqslant G\frac{\gamma_j(\hat{t})}{\gamma_i(t^\uparrow - 1) - \gamma_j(\hat{t})}. \tag{C.10}$$

Proof We start by proving Equation (C.8). Suppose $\gamma_j(\hat{t}) \geqslant \gamma_i(t^\uparrow - 1)$. Then:

$$\begin{aligned}\mu_j(\hat{t}) &\geqslant \mu_j(t^\uparrow - 1) + (\hat{t} - (t^\uparrow - 1))\gamma_j(\hat{t}) \\ &\geqslant \mu_i(t^\uparrow - 1) + (\hat{t} - (t^\uparrow - 1))\gamma_i(t^\uparrow - 1) \\ &\geqslant \mu_i(\hat{t})\end{aligned}$$

which is a contradiction with the definition of $\hat{t}$. Thus it must be $\gamma_j(\hat{t}) < \gamma_i(t^\uparrow - 1)$. Analogously,

suppose $\gamma_j(\hat{t}) < \gamma_i(t^{\downarrow})$. Then:

$$\begin{aligned}\mu_j(t^{\downarrow}) &\leqslant \mu_j(\hat{t}) + (t^{\downarrow} - \hat{t})\gamma_j(\hat{t}) \\ &< \mu_i(\hat{t}) - G + (t^{\downarrow} - \hat{t})\gamma_i(t^{\downarrow}) \\ &\leqslant \mu_i(t^{\downarrow}) - G\end{aligned}$$

which is a contradiction with the definition of $t^{\downarrow}$. Thus it must be $\gamma_j(\hat{t}) \geqslant \gamma_i(t^{\downarrow})$. We now prove Equation (C.9):

$$\begin{aligned}G &\leqslant \mu_i(\hat{t}) - \mu_j(\hat{t}) \leqslant \mu_i(t^{\uparrow} - 1) + (\hat{t} - (t^{\uparrow} - 1))\gamma_i(t^{\uparrow} - 1) \\ &- \mu_j(t^{\uparrow} - 1) - (\hat{t} - (t^{\uparrow} - 1))\gamma_j(\hat{t}) \\ &\leqslant (\hat{t} - (t^{\uparrow} - 1))(\gamma_i(t^{\uparrow} - 1) - \gamma_j(\hat{t}))\end{aligned}$$

and thus

$$\hat{t} - (t^{\uparrow} - 1) \geqslant G\frac{1}{\gamma_i(t^{\uparrow} - 1) - \gamma_j(\hat{t})}.$$

Finally, we prove Equation (C.10):

$$\begin{aligned}\mu_i(\hat{t}) - \mu_i(t^{\uparrow} - 1) &\geqslant \mu_j(\hat{t}) - \mu_j(t^{\uparrow} - 1) \geqslant (\hat{t} - (t^{\uparrow} - 1))\gamma_j(\hat{t}) \\ &\geqslant G\frac{\gamma_j(\hat{t})}{\gamma_i(t^{\uparrow} - 1) - \gamma_j(\hat{t})}.\end{aligned}$$

■

Lemma 54. *Let $M \in \mathbb{N}_{\geqslant 1}$, $M \geqslant 2$, $m_1 > m_2 > \cdots > m_M > m_{M+1} > 0$, then:*

$$\sum_{i=1}^{M} \frac{1}{m_i - m_{i+1}} \geqslant \frac{M^2}{m_1 - m_{M+1}}, \tag{C.11}$$

$$\sum_{i=1}^{M} \frac{m_{i+1}}{m_i - m_{i+1}} \geqslant \frac{M}{\left(\frac{m_1}{m_{M+1}}\right)^{\frac{1}{M}} - 1}. \tag{C.12}$$

Proof We regard $m_1 > m_{M+1} > 0$ as fixed constants and study the functions

$$f(m_2, \dots, m_M) = \sum_{i=1}^{M} \frac{1}{m_i - m_{i+1}},$$

$$g(m_2, \dots, m_M) = \sum_{i=1}^{M} \frac{m_{i+1}}{m_i - m_{i+1}}$$

defined for $m_1 > m_2 > \cdots > m_M > m_{M+1}$. Observe that the functions are defined on an open set and their values tend to infinity when the input tends to the border of the domain. We show that they have only one stationary point, which then must be a minimum point. We start by proving Equation (C.11). Let $k \in \{2, \dots, M\}$:

$$\frac{df}{dm_k}(m_2, \dots, m_M) = \frac{1}{(m_{k-1} - m_k)^2} - \frac{1}{(m_k - m_{k+1})^2} = 0$$

if and only if

$$m_{k+1} = 2m_k - m_{k-1}.$$

The linear system above is equivalent to:

$$m_i = (i-1)m_2 - (i-2)m_1 \quad \text{for} \quad i \in \{3, \dots, M+1\}. \tag{C.13}$$

Thus $m_{M+1} = Mm_2 - (M-1)m_1$, and then

$$m_2 = \frac{(M-1)m_1 + m_{M+1}}{M}.$$

By plugging this result into Equation (C.13), we get the coordinates of the minimum point:

$$m_i^* := \frac{(M+1-i)m_1 + (i-1)m_{M+1}}{M} \quad \text{for} \quad i \in \{2, \dots, M\}.$$

Thus:

$$f(m_1, \dots, m_M) \geqslant f(m_1^*, \dots, m_M^*) = \frac{M^2}{m_1 - m_{M+1}}.$$

We now prove Equation (C.12) analogously:

$$\frac{dg}{dm_k}(m_2, \dots, m_M) = \frac{m_{k-1}}{(m_{k-1} - m_k)^2} - \frac{m_{k+1}}{(m_k - m_{k+1})^2} = 0$$

if and only if

$$m_{k+1} = \frac{m_k^2}{m_{k-1}}$$

if and only if

$$\ln m_{k+1} = 2 \ln m_k - \ln m_{k-1}.$$

Observe that we get the same linear system of the previous case, with the difference that the variables are now $\ln m_i$. Thus, the solution is:

$$\ln m_i = \frac{(M+1-i)\ln m_1 + (i-1)\ln m_{M+1}}{M}$$

and then

$$m_i^* := m_1^{\frac{M+1-i}{M}} m_{M+1}^{\frac{i-1}{M}} \quad \text{for} \quad i \in \{2, \dots, M\}.$$

Finally:

$$g(m_2, \dots, m_M) \geqslant g(m_2^*, \dots, m_M^*) = \frac{M}{\left(\frac{m_1}{m_{M+1}}\right)^{\frac{1}{M}} - 1}.$$

■

Lemma 55. *Let $G \in (0, 1]$, $T' \in \mathbb{N}_{\geqslant 1}$, $M \in \mathbb{N}_{\geqslant 1}$, $i, j \in [k]$ such that there exist rounds*

$$2 \leqslant t_1^{\uparrow} \leqslant \hat{t}_1 < t_1^{\downarrow} \leqslant t_2^{\uparrow} \leqslant \hat{t}_2 < t_2^{\downarrow} \leqslant \cdots \leqslant t_M^{\uparrow} \leqslant \hat{t}_M \leqslant T'$$

which satisfy

$$i \uparrow_{t_l^{\uparrow}} j, \mu_i(\hat{t}_l) \geqslant \mu_j(\hat{t}_l) + G \text{ for all } l \in [M],$$

$$j \uparrow_{t_l^{\downarrow}} i \text{ for all } l \in [M-1].$$

Then:

$$M \leqslant 4 \ln(3T'/G) G^{-\frac{1}{2}}.$$

Proof Observe that, since

$$\begin{aligned} \mu_i(\hat{t}_M) &\geqslant \mu_j(\hat{t}_M) + G \geqslant \mu_j(t_M^{\uparrow} - 1) + G \\ &\geqslant \mu_i(t_M^{\uparrow} - 1) + G, \end{aligned}$$

we have

$$T' \gamma_i(t_M^{\uparrow} - 1) \geqslant (\hat{t}_M - (t_M^{\uparrow} - 1)) \gamma_i(t_M^{\uparrow} - 1) \geqslant \mu_i(\hat{t}_M) - \mu_i(t_M^{\uparrow} - 1) \geqslant G$$

and thus

$$\gamma_i(t_M^{\uparrow} - 1) \geqslant \frac{G}{T'}.$$

Now, assume $M \geqslant 3$. Then:

$$
\begin{aligned}
1 \geqslant \mu_i(T') - \mu_i(1) &\geqslant \sum_{l=1}^{M-1} (\mu_i(\hat{t}_l) - \mu_i(t_l^{\uparrow} - 1)) \\
&\geqslant G \sum_{l=1}^{M-1} \frac{\gamma_j(\hat{t}_l)}{\gamma_i(t_l^{\uparrow} - 1) - \gamma_j(\hat{t}_l)} && \text{(C.14)} \\
&\geqslant G \sum_{l=1}^{M-1} \frac{\gamma_i(t_{l+1}^{\uparrow} - 1)}{\gamma_i(t_l^{\uparrow} - 1) - \gamma_i(t_{l+1}^{\uparrow} - 1)} && \text{(C.15)} \\
&\geqslant G \frac{M-1}{\left(\frac{\gamma_i(t_1^{\uparrow}-1)}{\gamma_i(t_M^{\uparrow}-1)}\right)^{\frac{1}{M-1}} - 1} && \text{(C.16)} \\
&\geqslant G \frac{M-1}{\left(\frac{T'}{G}\right)^{\frac{1}{M-1}} - 1} = G \frac{M-1}{\exp\left(\frac{\ln(T'/G)}{M-1}\right) - 1} && \text{(C.17)}
\end{aligned}
$$

where line (C.14) follows from Lemma 53, line (C.15) follows from the fact that $\frac{x}{a-x}$ is non-decreasing for $a \geqslant 0$ and the concavity, line (C.16) follows from Lemma 54, and line (C.17) follows from the fact that $\gamma_i(t_1^{\uparrow} - 1) \leqslant 1$ and $\gamma_i(t_M^{\uparrow} - 1) \geqslant \frac{G}{T'}$. Now, if $M \geqslant 1 + \ln(T'/G)$, by Lemma 58, we have $\exp\left(\frac{\ln(T'/G)}{M-1}\right) - 1 \leqslant 3\frac{\ln(T'/G)}{M-1}$, and thus, by the chain of inequalities above:

$$
1 \geqslant G \frac{(M-1)^2}{3\ln(T'/G)} \text{ iff } M \leqslant 1 + \sqrt{3\ln(T'/G)G^{-1}}.
$$

Thus, by considering all possible cases, we have:

$$
M \leqslant \max\{2, \ln(T'/G), 1 + \sqrt{3\ln(T'/G)G^{-1}}\} \leqslant 4\ln(3T'/G)G^{-\frac{1}{2}}.
$$

■

We are now ready to prove Lemma 22.

Lemma 22. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $T \in \mathbb{N}_{\geqslant 1}$, $d \in (0, k]$, we have that:*

$$
|\mathcal{W}_{>d}(T)| \leqslant 9\ln\left(3e^{(\alpha)}_{w^{(\alpha)}(T)} k/d\right) k^{\frac{5}{2}} d^{-\frac{1}{2}}.
$$

Proof Let $w \in \mathcal{W}_{>d}(T)$. Then there exists $i \in [k]$ such that $d_w(i) > d$. But, in virtue of

Lemma 52, we have:

$$(|[i]_{w\times^+}| - 1) \max_{\substack{j,k\in[i]_{w\times^+} \text{ s.t.} j\ {}_w\times k \\ t\in\{s_w^{(\alpha)},\dots,e_w^{(\alpha)}\}}} |\mu_j(t) - \mu_k(t)| \geqslant d_w(i) > d.$$

Thus, there must be $j, k \in [i]_{w\times^+}$ and $t \in \{s_w^{(\alpha)}, \dots, e_w^{(\alpha)}\}$ such that $j\ {}_w\times k$ and

$$|\mu_j(t) - \mu_k(t)| > \frac{d}{|[i]_{w\times^+}| - 1} > \frac{d}{k}.$$

Observe that it must be either $i \times_{t'} j$ for $t' \leqslant t$ or $i \times_{t'} j$ for $t' > t$, with $t' \in \{s_w^{(\alpha)} + 1, \dots, e_w^{(\alpha)}\}$. W.l.o.g. we assume that i overtakes j. In the first case, window w must contain one of the rounds in which i overtakes j and then their expected rewards diverge by at least d/k. In the second case, window w must contain either the first round in which i overtakes j and which is right after one of the rounds in which i overtakes j and their expected rewards diverge by at least d/k or the first time in which i overtakes j. In virtue of Lemma 55 with $G = \frac{d}{k}$ and $T' = e^{(\alpha)}_{w^{(\alpha)}(T)}$, the rounds described in the first case are in number no more than $4\ln(3e^{(\alpha)}_{w^{(\alpha)}(T)}k/d)(d/k)^{-1/2}$, while the rounds described in the second case are in number no more than $4\ln(3e^{(\alpha)}_{w^{(\alpha)}(T)}k/d)(d/k)^{-1/2} + 1$ for a fixed choice of $i, j \in [k]$. Since we have at most k^2 such choices, it must be:

$$\begin{aligned}|\mathcal{W}_{>d}(T)| &\leqslant k^2(8\ln(3e^{(\alpha)}_{w^{(\alpha)}(T)}k/d)(d/k)^{-\frac{1}{2}} + 1) \\ &\leqslant 9\ln(3e^{(\alpha)}_{w^{(\alpha)}(T)}k/d)k^{\frac{5}{2}}d^{-\frac{1}{2}}.\end{aligned}$$

■

C.2.6. Proof of Theorem 23

The goal of this section is to prove Theorem 23. We start with an upper bound to $w^{(\alpha)}(T)$, $e^{(\alpha)}_{w^{(\alpha)}(T)}$, and $\Upsilon_{\boldsymbol{\nu}}\left(1, e^{(\alpha)}_{w^{(\alpha)}(T)}\right)$.

Lemma 56. *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $T \in \mathbb{N}_{\geqslant 2}$, we have:*

$$w^{(\alpha)}(T) \leqslant (2(1+\alpha)T)^{1/(1+\alpha)} \leqslant 4\alpha T^{1/(1+\alpha)}, \tag{C.18}$$

$$e^{(\alpha)}_{w^{(\alpha)}(T)} \leqslant 4(1+\alpha)T \leqslant 8\alpha T, \tag{C.19}$$

$$\Upsilon_{\boldsymbol{\nu}}\left(1, e^{(\alpha)}_{w^{(\alpha)}(T)}\right) \leqslant 8(1+\alpha)\Upsilon_{\boldsymbol{\nu}}(1,T) \leqslant 16\alpha\Upsilon_{\boldsymbol{\nu}}(1,T). \tag{C.20}$$

Proof We start by proving Equation (C.18). If $w \in \mathbb{N}_{\geqslant 1}$, $w \geqslant (2(1+\alpha)T)^{1/(1+\alpha)}$, then, by Lemma 50, we have:

$$e_w^{(\alpha)} \geqslant \frac{w^{1+\alpha}}{2(1+\alpha)} \geqslant T.$$

Thus it must be $w^{(\alpha)}(T) \leqslant (2(1+\alpha)T)^{1/(1+\alpha)}$. We now use Equation (C.18) to prove Equation (C.19).

$$\begin{aligned} e^{(\alpha)}_{w^{(\alpha)}(T)} &\leqslant w^{(\alpha)}(T) D^{(\alpha)}_{w^{(\alpha)}(T)} \\ &\leqslant 2(2(1+\alpha)T)^{\frac{1}{1+\alpha}} (2(1+\alpha)T)^{\frac{\alpha}{1+\alpha}} = 4(1+\alpha)T, \end{aligned} \tag{C.21}$$

where in line (C.21) we use the definition of $D_w^{(\alpha)}$, Equation (C.18), and the fact that $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$. Finally, we prove Equation (C.20).

$$\Upsilon_{\boldsymbol{\nu}}\left(1, e^{(\alpha)}_{w^{(\alpha)}(T)}\right) \leqslant \frac{e^{(\alpha)}_{w^{(\alpha)}(T)} - 1}{T-1} \Upsilon_{\boldsymbol{\nu}}(1,T) \tag{C.22}$$

$$\leqslant 2 \frac{e^{(\alpha)}_{w^{(\alpha)}(T)}}{T} \Upsilon_{\boldsymbol{\nu}}(1,T) \leqslant 8(1+\alpha)\Upsilon_{\boldsymbol{\nu}}(1,T), \tag{C.23}$$

where line (C.22) follows by applying iteratively Lemma 51 and line (C.23) follows from the fact that $T \geqslant 2$ and by Equation (C.19). ■

We are ready to prove Theorem 23.

Theorem 23 (Upper Bound for the Rising Concave Setting)**.** *For all restless rising concave MABs $\boldsymbol{\nu}$, $\alpha \geqslant 1$, $T \in \mathbb{N}_{\geqslant 24}$, we have that:*

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), T) \leqslant 2^{15}\alpha^3(1+\sigma)\left(\ln\left(\alpha k T^3\right)\right)^{\frac{3}{2}}\left(k^3 T^{\frac{3/4\alpha}{1+\alpha}} + k^3 T^{\frac{5/4\alpha-1}{1+\alpha}}\Upsilon_{\boldsymbol{\nu}}(1,T) + kT^{\frac{1+\alpha/2}{1+\alpha}}\right).$$

In particular, for $\alpha' := 8/3$, we get:

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha'), T) = \widetilde{\mathcal{O}}\left(k^3 T^{\frac{6}{11}} + k^3 T^{\frac{7}{11}}\Upsilon_{\boldsymbol{\nu}}(1,T) + kT^{\frac{7}{11}}\right).$$

Furthermore, for $\alpha'' := (8 - 8\log_T(k\sqrt{V_T}))/(3 + 8\log_T(k\sqrt{V_T}))$, under the additional assumptions $\boldsymbol{\nu} \in \mathcal{E}_{\mathrm{c}}(T, V_T)$, $T \geqslant \max\{k^{-8/3}V_T^{-4/3} + 1, k^{16/5}V_T^{8/5}\}$, we get:

$$R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha''), T) = \widetilde{\mathcal{O}}\left(k^{\frac{27}{11}} T^{\frac{6}{11}} V_T^{-\frac{3}{11}} + k^{\frac{15}{11}} T^{\frac{7}{11}} V_T^{\frac{2}{11}}\right).$$

Proof Let $d' \coloneqq kT^{-(\alpha/2)/(1+\alpha)} \in (0, k]$. Then:

$$
\begin{aligned}
R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \mathcal{W}_{>d'}(T)) &\leqslant |\mathcal{W}_{>d'}(T)| \max_{w \in \mathcal{W}_{>d'}(T)} \{3kB_w^{(\alpha)} + D_w^{(\alpha)} d_w^*\} && \text{(C.24)} \\
&\leqslant 9 \ln(3e^{(\alpha)}_{w^{(\alpha)}(T)} T^{\frac{\alpha/2}{1+\alpha}}) k^2 T^{\frac{\alpha/4}{1+\alpha}} && \text{(C.25)} \\
&\cdot \max_{w \in \mathcal{W}_{>d'}(T)} \{3kB_w^{(\alpha)} + 16\alpha k D_w^{(\alpha)} w^{-1} \Upsilon_{\boldsymbol{\nu}}(1, e_w^{(\alpha)})\} \\
&\leqslant 9 \ln(24\alpha T^2) k^2 T^{\frac{\alpha/4}{1+\alpha}} && \text{(C.26)} \\
&\cdot \max_{w \in \mathcal{W}_{>d'}(T)} \{6k(1 + 2\sigma\sqrt{D_w^{(\alpha)} \ln(2kD_w^{(\alpha)})}) \\
&+ 32\alpha k w^{\alpha-1} \Upsilon_{\boldsymbol{\nu}}(1, e_w^{(\alpha)})\} \\
&\leqslant 9 \ln(24\alpha T^2) k^2 T^{\frac{\alpha/4}{1+\alpha}} && \text{(C.27)} \\
&\cdot (6k(1 + 2\sigma\sqrt{8\alpha T^{\frac{\alpha}{1+\alpha}} \ln(16\alpha kT)}) \\
&+ 2^{11}\alpha^3 k T^{\frac{\alpha-1}{1+\alpha}} \Upsilon_{\boldsymbol{\nu}}(1, T)) \\
&\leqslant 2^4 \ln(\alpha k T^3) k^2 T^{\frac{\alpha/4}{1+\alpha}} && \text{(C.28)} \\
&\cdot 2^{11}\alpha^3(1+\sigma)(\ln(\alpha kT^3))^{\frac{1}{2}} k (T^{\frac{\alpha/2}{1+\alpha}} + T^{\frac{\alpha-1}{1+\alpha}} \Upsilon_{\boldsymbol{\nu}}(1, T)) \\
&= 2^{15}\alpha^3(1+\sigma)(\ln(\alpha kT^3))^{\frac{3}{2}} k^3 (T^{\frac{3/4\alpha}{1+\alpha}} + T^{\frac{5/4\alpha-1}{1+\alpha}} \Upsilon_{\boldsymbol{\nu}}(1, T))
\end{aligned}
$$

where line (C.24) follows from Lemma 20, line (C.25) follows from Lemma 22 and Lemma 21, line (C.26) follows from Lemma 56, the definition of $D_w^{(\alpha)}$, the fact that $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$, and the definition of $B_w^{(\alpha)}$, line (C.27) follows from the fact that the expression inside $\max$ is increasing in w, Lemma 56, and the fact that $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$, and line (C.28) follows from

$T \geqslant 24$. Furthermore:

$$
\begin{aligned}
R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \mathcal{W}_{\leqslant d'}(T)) &\leqslant |\mathcal{W}_{\leqslant d'}(T)| \max_{w \in \mathcal{W}_{\leqslant d'}(T)} \{3kB_w^{(\alpha)} + D_w^{(\alpha)} d_w^*\} \\
&\leqslant w^{(\alpha)}(T)(6k(1 + 2\sigma\sqrt{D^{(\alpha)}_{w^{(\alpha)}(T)} \ln(2kD^{(\alpha)}_{w^{(\alpha)}(T)})}) + D^{(\alpha)}_{w^{(\alpha)}(T)} d') \\
&\leqslant 4\alpha T^{\frac{1}{1+\alpha}}(1+\sigma)(\ln(\alpha k T^3))^{\frac{1}{2}}(12k\sqrt{8\alpha T^{\frac{\alpha}{1+\alpha}}} \\
&+ 8\alpha k T^{\frac{\alpha}{1+\alpha}} T^{-\frac{\alpha/2}{1+\alpha}}) \\
&\leqslant 2^9 \alpha^2 (1+\sigma)(\ln(\alpha k T^3))^{\frac{3}{2}} k T^{\frac{1+\alpha/2}{1+\alpha}} \\
&\leqslant 2^{15} \alpha^3 (1+\sigma)(\ln(\alpha k T^3))^{\frac{3}{2}} k T^{\frac{1+\alpha/2}{1+\alpha}}
\end{aligned}
$$

(C.29)

(C.30)

(C.31)

where line (C.29) follows from Lemma 20, line (C.30) follows from the definitions of $B_w^{(\alpha)}$ and $\mathcal{W}_{\leqslant d'}(T)$, and line (C.31) follows from Lemma 56, $T \geqslant 24$, $\lceil x \rceil \leqslant 2x$ for $x \geqslant 1$, and the definition of d'. By summing the previous results:

$$
\begin{aligned}
R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), T) &\leqslant R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \mathcal{W}_{>d'}(T)) + R_{\boldsymbol{\nu}}(\texttt{RC-BE}(\alpha), \mathcal{W}_{\leqslant d'}(T)) \\
&\leqslant 2^{15} \alpha^3 (1+\sigma)(\ln(\alpha k T^3))^{\frac{3}{2}} (k^3 T^{\frac{3/4\alpha}{1+\alpha}} + k^3 T^{\frac{5/4\alpha - 1}{1+\alpha}} \Upsilon_{\boldsymbol{\nu}}(1, T) \\
&+ k T^{\frac{1+\alpha/2}{1+\alpha}}).
\end{aligned}
$$

Finally, observe that, under the additional assumption $\boldsymbol{\nu} \in \mathcal{E}_{\text{c}}(T, V_T)$, we have $\Upsilon_{\boldsymbol{\nu}}(1, T) \leqslant V_T$, and $T \geqslant \max\{k^{-8/3} V_T^{-4/3} + 1, k^{16/5} V_T^{8/5}\}$ guarantees $\alpha'' \geqslant 1$. ■

C.3. Technical Lemmas

Lemma 57. *Let $\delta \in \left(0, \frac{1}{2}\right)$. Let $x_1, x_2 \in [\delta, 1-\delta]$, $x_1 \leqslant x_2$. Then:*

$$
\mathrm{D}_{\mathrm{KL}}(x_1 \| x_2) \leqslant \frac{(x_2 - x_1)^2}{2\ln(2)\delta^2}
$$

where $\mathrm{D}_{\mathrm{KL}}(\cdot\|\cdot)$ is defined as in Appendix C.1.

Proof Consider the function:

$$f(y) = \mathrm{D}_{\mathrm{KL}}(x_1 \| x_1 + y) \text{ for } y \in [0, x_2 - x_1].$$

Then $f(0) = 0$, $f'(0) = 0$ and

$$\begin{aligned} f''(y) &= \frac{1}{\ln(2)} \left(\frac{x_1}{(x_1 + y)^2} + \frac{1 - x_1}{(y + x_1 - 1)^2} \right) \leqslant \frac{1}{\ln(2)} \left(\frac{1}{x_1} + \frac{1}{1 - x_1} \right) \\ &= \frac{1}{\ln(2) x_1 (1 - x_1)} \leqslant \frac{1}{\ln(2)\delta^2}. \end{aligned}$$

Thus:

$$f(y) = f(0) + \int_0^y \left(f'(0) + \int_0^{y_1} f''(y_1) dy_1 \right) dy \leqslant \frac{y^2}{2 \ln(2) \delta^2}.$$

The result follows from the fact that $\mathrm{D}_{\mathrm{KL}}(x_1 \| x_2) = f(x_2 - x_1)$. ■

Lemma 58.

$$e^x - 1 \leqslant 3x \quad \textit{for} \quad x \in [0, 1].$$

Proof Let $f(x) = e^x - 1$. Then: $f'(x) = e^x = f''(x)$. Thus, by Taylor's theorem, if $x \in [0, 1]$, there exists $\xi \in (0, 1)$ such that

$$f(x) = f(0) + f'(0)x + \frac{f''(\xi)}{2} x^2 = x \left(1 + \frac{e^\xi}{2} x \right) \leqslant x \left(1 + \frac{e}{2} \right) \leqslant 3x.$$

■

C.4. Numerical Simulations

In this appendix, we present additional numerical simulations which compare `RC-BE`(α) with the baseline algorithms reported in Section 3.3.6. Furthermore, we provide information regarding the compute resources used to run the simulations.

Baselines. We consider the following baseline algorithms:

- `Rexp3` (Besbes et al., 2014), an algorithm for restless MABs based on a variation budget for the expected rewards of the arms over the learning horizon.
- `R-less-UCB` (Metelli et al., 2022), an algorithm for restless rising concave MABs which relies on the optimism principle and exploits the structure of the setting through a specifically

crafted estimator.

- `UCB1` (Auer et al., 2002a), one of the most effective algorithms for stationary MABs.

The choices of the parameters of the algorithms that we compared are the following:

- `Rexp3`: $V_T = k$ since, as remarked in Section 3.3.2, in the rising setting the cumulative increment is always smaller than or equal to k; $D_T = \lceil (k \ln(k))^{1/3} (T/V_T)^{2/3} \rceil$; $\gamma = \min \left\{ 1, \sqrt{k \ln(k)/(D_T(e-1))} \right\}$ as recommended in (Besbes et al., 2014).
- `R-less-UCB`: $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ where $N_{i,t-1}$ is the number of times arm i has been pulled by the agent in the first $t-1$ rounds, with $\epsilon \in (0, 1/2)$; $\alpha > 2$ as prescribed in (Metelli et al., 2022). In particular, we choose $\epsilon = 0.25$; $\alpha = 2.1$.
- `UCB1`: the upper confidence bound interval for arm i at round t is $\sigma \sqrt{4 \ln(t)/N_{i,t-1}}$.

C.5. Flaw in the Original Analysis of k-armed `Budgeted Exploration`

In this appendix, we highlight a flaw in the original analysis of the extension of `Budgeted Exploration` in the k-armed setting, which is presented in the unpublished preprint (Jia et al., 2024). For notation and definitions, refer to the original paper. The analysis relies on the following proposition, stated in Lemma I.7: "First, we observe that on the clean event C, any arm in A^* can never be eliminated for "losing" to an arm in $(A^*)^c$". It is possible to construct a counterexample that satisfies the hypotheses of the lemma and violates the previous proposition. We now show how. We work with 3 arms. We describe the evolution of the expected reward of the arms only in a certain window. This is sufficient for the construction of the counterexample since the lemma regards the behavior of the algorithm in a single window. The window is composed of $17W$ rounds, with $W \in \mathbb{N}_{\geqslant 2}$ to be chosen later. The expected rewards of the arms are defined as follows:

$$r_a(t) = f_a\left(\frac{t}{17W}\right) \text{ for } t \in [17W]$$

where $f_a : [0,1] \to [-1,1]$ is a 2-Hölder function with Lipschitz constant $L > 0$ for $a \in [3]$. More specifically, we choose:

- The function in which the expected rewards of the first arm are embedded as:

$$f_1(x) = \frac{1}{2} - \int_0^x \begin{cases} 0 & \text{if } t \in \left[0, \frac{3}{17} + 2\sqrt{\frac{d}{L}}\right] \\ L\left(t - \frac{3}{17} - 2\sqrt{\frac{d}{L}}\right) & \text{if } t \in \left(\frac{3}{17} + 2\sqrt{\frac{d}{L}}, \frac{3}{17} + 3\sqrt{\frac{d}{L}}\right] \\ \sqrt{dL} - L\left(t - \frac{3}{17} - 3\sqrt{\frac{d}{L}}\right) & \text{if } t \in \left(\frac{3}{17} + 3\sqrt{\frac{d}{L}}, \frac{3}{17} + 4\sqrt{\frac{d}{L}}\right] \\ 0 & \text{if } t \in \left(\frac{3}{17} + 4\sqrt{\frac{d}{L}}, 1\right] \end{cases} dt.$$

- The function in which the expected rewards of the second arm are embedded as:

$$f_2(x) = \frac{1}{2} - \varepsilon.$$

- The function in which the expected rewards of the third arm are embedded as:

$$f_3(x) = \frac{1}{2} - d + \int_0^x \begin{cases} 0 & \text{if } t \in \left[0, \frac{3}{17}\right] \\ L\left(t - \frac{3}{17}\right) & \text{if } t \in \left(\frac{3}{17}, \frac{3}{17} + \sqrt{\frac{d}{L}}\right] \\ \sqrt{dL} - L\left(t - \frac{3}{17} - \sqrt{\frac{d}{L}}\right) & \text{if } t \in \left(\frac{3}{17} + \sqrt{\frac{d}{L}}, \frac{3}{17} + 2\sqrt{\frac{d}{L}}\right] \\ 0 & \text{if } t \in \left(\frac{3}{17} + 2\sqrt{\frac{d}{L}}, 1\right] \end{cases} dt.$$

The definitions rely on the constants $d, \varepsilon > 0$, $\varepsilon < d \leqslant 1/2$, which we choose later. To guarantee that the functions are well-defined, we impose:

$$4\sqrt{\frac{d}{L}} \leqslant \frac{2}{17} \quad \text{iff} \quad L \geqslant 34^2 d. \tag{C.32}$$

We work with deterministic rewards, which can be regarded as a special realization under the clean event C. Let $Z_a^{\text{total},t}$ be the cumulative reward of arm $a \in [k]$ observed up to round $t \in [17W]$, included. Assuming there is no elimination before round $3W$ (we choose d and ε in such a way that this is true), we have that:

$$Z_1^{\text{total},3W} = \frac{1}{2}W, \quad Z_2^{\text{total},3W} = \left(\frac{1}{2} - \varepsilon\right)W, \quad Z_3^{\text{total},3W} = \left(\frac{1}{2} - d\right)W.$$

Then:

$$Z_1^{\text{total},3W} - Z_2^{\text{total},3W} = \varepsilon W, \quad Z_1^{\text{total},3W} - Z_3^{\text{total},3W} = dW.$$

Let:

$$d := \frac{B}{W-1}, \quad \varepsilon := \frac{B}{2W}$$

where B is the budget of the algorithm. These choices are such that we eliminate arm 3 at the end of round $3W$ (and not before), losing to arm 1. Arm 2, instead, stays alive. To satisfy $d \leqslant 1/2$, it is sufficient to require $W \geqslant 3B$. After round $3W$, the algorithm pulls only arms 1 and 2. When $r_1(t) \geqslant r_2(t)$, their difference is at most ε. Thus:

$$Z_1^{\text{total},5W} - Z_2^{\text{total},5W} \leqslant 2\varepsilon W = B.$$

Hence, arm 2 is not eliminated before round $5W$ (included). By the choice of the instance, in virtue of Equation (C.32), after round $5W$, we have $r_1(t) = 1/2 - d$. Thus, after each round robin cycle, which takes 2 rounds, $Z_2^{\text{total},t} - Z_1^{\text{total},t}$ increases by $d - \varepsilon$. Then:

$$Z_2^{\text{total},17W} - Z_1^{\text{total},17W} = 6(d-\varepsilon)W - (Z_1^{\text{total},5W} - Z_2^{\text{total},5W}) \geqslant 3B - B = 2B.$$

This means that, at some point after round $5W$, arm 1 will be eliminated, losing to arm 2. But it is evident that $1 \in A^*$ and $2 \in (A^*)^c$. However, it is important to notice that $2 \in \mathcal{I}_w^\times$, consistent with our analysis. It remains to show that there are choices of B, W, T, and L which satisfy the hypotheses of the lemma and the additional requirements we imposed. In particular, they need to satisfy:

$$\begin{cases} \sqrt{\frac{17W\ln(3)\ln(T)}{3}} \leqslant B \leqslant \frac{W}{3} \\ 34^2 \frac{B}{W-1} \leqslant L \\ 17W \leqslant T \\ 2 \leqslant W \end{cases}.$$

It is clear that such an assignment exists. Furthermore, we can find such an assignment even when we restrict the budget to the natural choice, which has order $W^{1/2}$.

D | Adaptation to Unknown Distributional Parameters in Heavy-Tailed Bandits

D.1. Additional Related Works

In this section, we provide additional related works concerning adaptivity in statistics via Lepskii method and adaptivity in the case of subgaussian bandits.

D.1.1. Adaptivity via Lepskii Method

In Bhatt et al. (2022b), authors provide a novel technique to extend Catoni's M-estimator (Catoni, 2012) to the infinite variance setting. In principle, their procedure relies on the knowledge of both ϵ and the centered moment v, however, they propose a strategy based on the Lepskii method (Lepskii, 1992) to adapt to unknown v. While the Lepskii method is a popular choice in the adaptive statistics literature, we point out how it requires an upper bound on the quantity to estimate. Indeed, this method can be safely applied when adapting to unknown ϵ (since it can be at most 1), but when it comes to u (or the centered moment v), requiring an upper bound makes the approach *not* fully adaptive.

D.1.2. Adaptivity in Subgaussian Bandits

In the literature of subgaussian stochastic bandits, σ (the subgaussian proxy) is usually assumed to be known by the agent. However, many works consider settings in which this quantity is unknown. In this section, we discuss standard approaches to adapt to σ (or estimate it) in subgaussian bandits, and show the additional difficulties implied by the heavy-tailed setting.

The main difference between σ and u is that the former can be estimated from data while guaranteeing strong convergence properties. In Audibert et al. (2009), authors introduce `UCB-V`, a variation of the well-known `UCB1` algorithm capable of using a data-driven estimation of the

variance while keeping optimal performance. As customary in most of the literature, rewards are assumed to be bounded in a known range. However, in heavy-tailed bandits, it is not possible to make such an assumption, and the estimation of u cannot be carried on. Other works try to relax the assumption of bounded rewards by the means of other assumptions, *e.g.,* a known upper bound on kurtosis (Lattimore, 2017), or Gaussian rewards (Cowan et al., 2018).

Without additional assumptions, dealing with both the unknown range of the rewards and unknown σ comes at a cost. As shown in Hadiji and Stoltz (2023), when the range of the rewards is unknown and no additional knowledge on the distributions is available, it is impossible to be simultaneously optimal in both the instance-dependent sense and the worst-case one. The existence of such a trade-off shows how difficult is, even in subgaussian bandits, to attain optimal performances when no knowledge is given on the environment. As a consequence, also in fully adaptive heavy-tailed bandits, such an impossibility result holds. However, as we have discussed, thanks to a specific assumption not involving ϵ nor u we can provide optimal regret guarantees in both cases.

D.2. Proofs and Derivations

In this section, we prove the main theoretical results outlined in the paper.

D.2.1. Lower Bounds

Theorem 25 (Minimax lower bound – u-adaptive)**.** *Fix $\epsilon \in (0, 1]$. For every algorithm π, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $k \in \mathbb{N}_{\geqslant 2}$, it holds that:*

$$\sup_{u \geqslant 0} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^k} \frac{R_{\boldsymbol{\nu},T}(\pi)}{u^{\frac{1}{1+\epsilon}}} = +\infty. \tag{4.4}$$

More precisely, for every $u' \geqslant u \geqslant 0$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon, u')$ such that:

$$\max\left\{ \frac{R_{\boldsymbol{\nu},T}(\pi)}{u^{\frac{1}{1+\epsilon}}}, \frac{R_{\boldsymbol{\nu}',T}(\pi)}{(u')^{\frac{1}{1+\epsilon}}} \right\} \geqslant c_1 \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(1+\epsilon)^2}} T^{\frac{1}{1+\epsilon}}, \tag{4.5}$$

where $c_1 > 0$ is a constant independent of u, u', and T.

Proof We start by constructing two heavy-tailed bandit instances with a common maximum order of moment ϵ, but where $u' \geqslant u$. We use δ_x to denote the Dirac delta distribution centered on x

Base instance

$$\boldsymbol{\nu} = \begin{cases} \nu_1 = \delta_0, \\ \nu_2 = \left(1 - \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}\right) \delta_0 + \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \delta_{u^{\frac{1}{\epsilon}} \Delta^{-\frac{1}{\epsilon}}}, \end{cases} \tag{D.1}$$

where $\Delta \in (0, u^{\frac{1}{1+\epsilon}})$. Thus, we have $\mu_1 = 0$ and $\mu_2 = \Delta$. Furthermore, $\mathbb{E}_{X \sim \nu_1}[|X|^{1+\epsilon}] = 0$ and $\mathbb{E}_{X \sim \nu_2}[|X|^{1+\epsilon}] = u$ Therefore, the optimal arm is arm 2 and $\boldsymbol{\nu} \in \mathcal{P}(\epsilon, u)^2$.

Alternative instance

$$\boldsymbol{\nu}' = \begin{cases} \nu'_1 = \left(1 - (2\Delta)^{1+\frac{1}{\epsilon}} (u')^{-\frac{1}{\epsilon}}\right) \delta_0 + (2\Delta)^{1+\frac{1}{\epsilon}} (u')^{-\frac{1}{\epsilon}} \delta_{(u')^{\frac{1}{\epsilon}} (2\Delta)^{-\frac{1}{\epsilon}}}, \\ \nu'_2 = \nu_2, \end{cases} \tag{D.2}$$

where $\Delta \in (0, \frac{1}{2}(u')^{\frac{1}{1+\epsilon}})$. Thus we have $\mu'_1 = 2\Delta$ and $\mu'_2 = \Delta$. Furthermore, $\mathbb{E}_{X \sim \nu'_1}[|X|^{1+\epsilon}] = u'$ and $\mathbb{E}_{X \sim \nu'_2}[|X|^{1+\epsilon}] = u$. Therefore, the optimal arm is arm 1 and $\boldsymbol{\nu} \in \mathcal{P}(\epsilon, u')^2$.

We seek to prove that for any algorithm π, it holds that:

$$\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}}\right\}\geqslant f(T,\epsilon,u,u'),$$

being f a function increasing in T. The proof merges the approach of (Bubeck et al., 2013b, Theorem 5) with that of (Lattimore and Szepesvári, 2020, Chapters 14.2, 14.3).

First, we observe that:

$$\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}}\right\}\geqslant\frac{R_T(\pi,\boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}}=\frac{\Delta\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}},\tag{D.3}$$

where $\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]$ is the expected number of times arm 1 is pulled over the horizon T. Second, recalling which are the optimal arms in the two instances and that $u'\geqslant u$, we have:

$$\begin{aligned}
&\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}}\right\}\geqslant\\
&\qquad\geqslant(u'T)^{-\frac{1}{\epsilon+1}}\frac{\Delta T}{2}\max\left\{\mathbb{P}_{\pi,\boldsymbol{\nu}}\left(N_1(T)\geqslant T/2\right),\mathbb{P}_{\pi,\boldsymbol{\nu}'}\left(N_1(T)<T/2\right)\right\}\\
&\qquad\geqslant\frac{\Delta}{4}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\left(\mathbb{P}_{\pi,\boldsymbol{\nu}}\left(N_1(T)\geqslant T/2\right)+\mathbb{P}_{\pi,\boldsymbol{\nu}'}\left(N_1(T)<T/2\right)\right)\\
&\qquad\geqslant\frac{\Delta}{8}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\exp\left(-\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]D_{\mathrm{KL}}(\nu_1\|\nu_1')\right).
\end{aligned}\tag{D.4}$$

where we used Bretagnolle-Huber inequality and divergence decomposition, together with $\max\{a,b\}\geqslant\frac{1}{2}(a+b)$ for $a,b\geqslant 0$. Let us now compute the KL-divergence, noting that $\nu_1\ll\nu_1'$:

$$\begin{aligned}
D_{\mathrm{KL}}(\nu_1\|\nu_1')&=\nu_1(0)\ln\frac{\nu_1(0)}{\nu_1'(0)}\\
&=\ln\frac{1}{1-(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}}}\leqslant c(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}},
\end{aligned}\tag{D.5}$$

for $\Delta\in(0,\left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}}(u')^{\frac{1}{1+\epsilon}})$ and some constant $c\in(1,2)$. Putting together Equations (D.3),

(D.4) and (D.5), we have:

$$\begin{aligned}
\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} &\geqslant \\
&\geqslant \max\left\{\frac{\Delta\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}},\frac{\Delta}{8}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\exp\left(-c\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)](2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right)\right\} \\
&\geqslant \frac{\Delta}{2}\left(\frac{\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}}+\frac{1}{8}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\exp\left(-c\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)](2\Delta)^{\frac{1+\epsilon}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right)\right) \\
&\geqslant \frac{\Delta}{2}\min_{x\in[0,T]}\left\{\frac{x}{(uT)^{\frac{1}{1+\epsilon}}}+\frac{1}{8}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\exp\left(-cx(2\Delta)^{\frac{1+\epsilon}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right)\right\} =: g(x)
\end{aligned}$$

The latter is a convex function of x and the minimization can be carried out in closed form, vanishing the derivative and finding:

$$x^* = c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\ln\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}c(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right),$$

which leads to:

$$g(x^*) = \frac{\Delta}{2}(uT)^{-\frac{1}{\epsilon+1}}c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\ln\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}ec(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right).$$

We choose Δ such that:

$$\frac{Tu^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}c(2\Delta)^{\frac{1+\epsilon}{\epsilon}} = e^{\epsilon},$$

resulting in $\Delta = 2^{\frac{2\epsilon-1}{1+\epsilon}}e^{\frac{\epsilon^2}{1+\epsilon}}(cT)^{-\frac{\epsilon}{\epsilon+1}}u^{-\frac{\epsilon}{(\epsilon+1)^2}}(u')^{\frac{1+2\epsilon}{(\epsilon+1)^2}}$. This implies, after some calculations, that:

$$g(x^*) = c^{-\frac{\epsilon}{\epsilon+1}}2^{-\frac{2\epsilon+5}{\epsilon+1}}(1+\epsilon)e^{-\frac{\epsilon}{\epsilon+1}}u^{-\frac{\epsilon}{(\epsilon+1)^2}}(u')^{\frac{\epsilon}{(\epsilon+1)^2}} \geqslant c_1\left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}},$$

where $c_1 > 0$ is a value independent of T and both u and u'. Finally, we have that

$$\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geqslant c_1\left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}}.$$

We observe that $\Delta < \left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}}(u')^{\frac{1}{1+\epsilon}}$ for sufficiently large T. This concludes the proof of the second statement. For the first statement, we observe that, since $u' \geqslant u$ can be taken arbitrarily large, the right-hand side of this inequality can be arbitrarily large. ■

Theorem 26 (Minimax lower bound – ϵ-adaptive)**.** *Fix $u = 1$. For every algorithm π, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $k \in \mathbb{N}_{\geqslant 0}$, it holds that:*

$$\sup_{\epsilon \in (0,1]} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^k} \frac{R_{\boldsymbol{\nu},T}(\pi)}{T^{\frac{1}{1+\epsilon}}} \geqslant c_2 T^{\frac{1}{16}}. \tag{4.6}$$

More precisely, for every $\epsilon, \epsilon' \in (0,1]$ with $\epsilon' \leqslant \epsilon$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon', u)$ such that:

$$\max\left\{ \frac{R_{\boldsymbol{\nu},T}(\pi)}{T^{\frac{1}{1+\epsilon}}}, \frac{R_{\boldsymbol{\nu}',T}(\pi)}{T^{\frac{1}{1+\epsilon'}}} \right\} \geqslant c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon)(1+\epsilon')^2}}, \tag{4.7}$$

where $c_2 > 0$ is a constant independent of ϵ, ϵ', and T.

Proof We start by constructing two heavy-tailed bandit instances with different maximum orders of moment ϵ and ϵ', where $0 < \epsilon' < \epsilon < 1$. For the sake of simplicity, but without loss of generality, we will assume a common (and known to the algorithm) maximum moment of $u = 1$.

Base instance

$$\boldsymbol{\nu} = \begin{cases} \nu_1 = \delta_0, \\ \nu_2 = (1 + \Delta\gamma - \gamma^{1+\epsilon})\delta_0 + (\gamma^{1+\epsilon} - \Delta\gamma)\delta_{1/\gamma} \end{cases}, \tag{D.6}$$

where $\Delta \in [0, 1/2]$ and $\gamma = (2\Delta)^{\frac{1}{\epsilon}}$. Thus, we have $\mu_1 = 0$ and $\mu_2 = \Delta$. Furthermore, $\mathbb{E}_{X\sim\nu_1}[|X|^\alpha] = 0$ and $\mathbb{E}_{X\sim\nu_2}[|X|^\alpha] = 2^{\frac{1-\alpha}{\epsilon}} \Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}$, which are guaranteed to be bounded by a constant smaller than 1 only if $\alpha \leqslant \epsilon + 1$. Thus, this instance admits moments finite only up to order $\epsilon + 1$, *i.e.,* $\boldsymbol{\nu} \in \mathcal{P}(\epsilon, 1)^2$. Moreover, the optimal arm is arm 2.

Alternative instance

$$\boldsymbol{\nu}' = \begin{cases} \nu'_1 = (1 - (\gamma')^{1+\epsilon'})\delta_0 + (\gamma')^{1+\epsilon'}\delta_{1/\gamma'}, \\ \nu'_2 = \nu_2 \end{cases}, \tag{D.7}$$

where $\Delta \in [0, 1/2]$ and $\gamma' = (2\Delta)^{\frac{1}{\epsilon'}}$. Thus, we have $\mu'_1 = 2\Delta$ and $\mu'_2 = \Delta$. Furthermore, $\mathbb{E}_{X\sim\nu'_1}[|x|^\alpha] = (2\Delta)^{\frac{1+\epsilon'-\alpha}{\epsilon'}}$ and $\mathbb{E}_{X\sim\nu'_2}[|x|^\alpha] = 2^{\frac{1-\alpha}{\epsilon}} \Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}$, which are guaranteed to be bounded by a constant smaller than 1 only if $\alpha \leqslant \epsilon' + 1$. Thus, this instance admits moments finite only up to order $\epsilon' + 1$, *i.e.,* $\boldsymbol{\nu} \in \mathcal{P}(\epsilon', 1)^2$. Moreover, the optimal arm is arm 1.

We will prove, that for any algorithm π it holds that:

$$\max\left\{ \frac{R_T(\pi, \boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\pi, \boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}} \right\} \geqslant f(T, \epsilon, \epsilon'),$$

being f a function increasing in T. The proof emulates the analyses and steps performed to prove Theorem 25. First, we observe that:

$$\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\}\geqslant\frac{R_T(\pi,\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}=\frac{\Delta\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}},\tag{D.8}$$

where $\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]$ is the expected number of times arm 1 is pulled over the horizon T.

Second, recalling which are the optimal arms in the two instances and that $\epsilon'<\epsilon$, we have:

$$\begin{aligned}
&\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\}\geqslant\\
&\quad\geqslant T^{-\frac{1}{\epsilon'+1}}\max\left\{\frac{\Delta T}{2}\mathbb{P}_{\pi,\boldsymbol{\nu}}\left(N_1(T)\geqslant\frac{T}{2}\right),\frac{\Delta T}{2}\mathbb{P}_{\pi,\boldsymbol{\nu}'}\left(N_1(T)<\frac{T}{2}\right)\right\}\\
&\quad\geqslant\frac{\Delta}{4}T^{\frac{\epsilon'}{\epsilon'+1}}\left(\mathbb{P}_{\pi,\boldsymbol{\nu}}\left(N_1(T)\geqslant\frac{T}{2}\right)+\mathbb{P}_{\pi,\boldsymbol{\nu}'}\left(N_1(T)<\frac{T}{2}\right)\right)\\
&\quad\geqslant\frac{\Delta}{8}T^{\frac{\epsilon'}{\epsilon'+1}}\exp\left(-\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]D_{\mathrm{KL}}(\nu_1\|\nu_1')\right).
\end{aligned}\tag{D.9}$$

where we used Bretagnolle-Huber inequality and divergence decomposition, together with $\max\{a,b\}\geqslant\frac{1}{2}(a+b)$ for $a,b\geqslant 0$. Let us now compute the KL-divergence, noting that $\nu_1\ll\nu_1'$:

$$\begin{aligned}
D_{\mathrm{KL}}(\nu_1\|\nu_1')&=\nu_1(0)\ln\frac{\nu_1(0)}{\nu_1'(0)}\\
&=\ln\frac{1}{1-(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}}\leqslant c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}},
\end{aligned}\tag{D.10}$$

for $\Delta\in[0,1/4]$ and some constant $c\in(1,2)$. Putting together Equations (D.8), (D.9) and (D.10), we have:

$$\begin{aligned}
&\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}},\frac{R_T(\pi,\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\}\\
&\quad\geqslant\max\left\{\frac{\Delta\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}},\frac{\Delta}{8}T^{\frac{\epsilon'}{\epsilon'+1}}\exp\left(-c\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_1(T)](2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right\}\\
&\quad\geqslant\frac{\Delta}{2}\left(\frac{\mathbb{E}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}}+\frac{1}{8}T^{\frac{\epsilon'}{\epsilon'+1}}\exp\left(-c\mathbb{E}[N_1(T)](2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right)\\
&\quad\geqslant\frac{\Delta}{2}\min_{x\in[0,T]}\left\{\frac{x}{T^{\frac{1}{1+\epsilon}}}+\frac{1}{8}T^{\frac{\epsilon'}{\epsilon'+1}}\exp\left(-cx(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right\}=:g(x).
\end{aligned}$$

The latter is a convex function of x and the minimization can be carried out in closed form

vanishing the derivative and obtaining:

$$x^* = c^{-1}(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \ln\left(\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{8} c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right),$$

which leads to:

$$g(x^*) = \frac{\Delta}{2} T^{-\frac{1}{\epsilon+1}} c^{-1}(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \ln\left(\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{8} ec(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right).$$

We take Δ such that:

$$\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{8} c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}} = 1,$$

resulting in $\Delta = 2^{\frac{2\epsilon'-1}{1+\epsilon'}} c^{-\frac{\epsilon'}{1+\epsilon'}} T^{-\frac{\epsilon'}{1+\epsilon'}\left(\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}\right)}$. This imply, after some calculations, that:

$$g(x^*) = 2^{\frac{-2\epsilon'-5}{1+\epsilon'}} c^{-\frac{\epsilon'}{1+\epsilon'}} T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}} \geqslant c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}}.$$

where $c_2 > 0$ is a value independent of T and can be always selected to be ϵ and ϵ'. Finally, we have that:

$$\max\left\{\frac{R_T(\pi,\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\pi,\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\} \geqslant c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}}.$$

We observe that $\Delta < 1/4$ for sufficiently large T. We conclude by observing that the exponent of T is maximized by taking $\epsilon = 1$ and $\epsilon' = 1/3$. ■

Theorem 27 (Minimax lower bound under Assumption 6 - non-adaptive)**.** *Fix $\epsilon \in (0,1]$ and $u \geqslant 0$. For every algorithm π, sufficiently large learning horizon $T \in \mathbb{N}$, and every number of arms $k \in \mathbb{N}_{\geqslant 2}$, it holds that:*

$$\sup_{\substack{\boldsymbol{\nu}\in\mathcal{P}_{HT}(\epsilon,u)^k \\ \boldsymbol{\nu} \text{ fulfills Assumption 6}}} R_{\boldsymbol{\nu},T}(\pi) \geqslant c_3 k^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}}, \tag{4.8}$$

where $c_3 > 0$ is a constant independent of u, ϵ, k and T.

Proof We will construct instances using the following prototype of reward distribution, defined for $y \in (0, u^{\frac{1}{1+\epsilon}})$ and $\Delta \in (0, u^{\frac{1}{1+\epsilon}})$:

$$\rho_y = \left(1 - y^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}\right)\delta_0 + \left(y^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}\right)\delta_{-u^{\frac{1}{\epsilon}}\Delta^{-\frac{1}{\epsilon}}}. \tag{D.11}$$

The two instances are constructed by the means of Equation (D.11). Note that we have:

$$\mathop{\mathbb{E}}_{X\sim\rho_y}[X] = -y^{1+\frac{1}{\epsilon}}\Delta^{-\frac{1}{\epsilon}}, \tag{D.12}$$

$$\mathop{\mathbb{E}}_{X\sim\rho_y}[|X|^{1+\epsilon}] = y^{1+\frac{1}{\epsilon}}\Delta^{-1-\frac{1}{\epsilon}}u \leqslant u, \tag{D.13}$$

for every $0 \leqslant y \leqslant \Delta$.

Base instance

$$\boldsymbol{\nu} = \begin{cases} \nu_1 = \rho_{\left(\frac{2}{3}\right)^{\frac{\epsilon}{1+\epsilon}}\Delta}, \\ \nu_j = \rho_\Delta, & j \in [k]\backslash\{1\}. \end{cases}$$

Alternative instance

$$\boldsymbol{\nu}' = \begin{cases} \nu'_1 = \rho_{\left(\frac{2}{3}\right)^{\frac{\epsilon}{1+\epsilon}}\Delta}, \\ \nu'_i = \rho_{\left(\frac{1}{3}\right)^{\frac{\epsilon}{1+\epsilon}}\Delta}, \\ \nu'_j = \rho_\Delta, & j \in [k]\backslash\{1,i\}, \end{cases}$$

where $i \in \operatorname{argmin}_{j\neq 1}\mathbb{E}_{\pi,\boldsymbol{\nu}'}[N_j(T)]$. For the base instance, we have $\mu_1 = -2\Delta/3$ and $\mu_j = -\Delta$ for all $j \neq 1$; whereas for the alternative instance $\mu'_j = \mu_j$ for all $j \neq i$ and $\mu'_i = -\Delta/3$. Both instances satisfy Assumption 6, being the support a subset made of non-positive numbers. Moreover, for the base instance, the optimal arm is 1 and for the alternative instance, the optimal arm is i. Using the Bretagnolle-Huber inequality, we obtain:

$$\begin{aligned} R_T(\pi,\boldsymbol{\nu}) + R_T(\pi,\boldsymbol{\nu}') &\geqslant \frac{\Delta T}{6}\left(\mathbb{P}_{\pi,\boldsymbol{\nu}}\left(N_1 \leqslant \frac{T}{2}\right) + \mathbb{P}_{\pi,\boldsymbol{\nu}'}\left(N_1 > \frac{T}{2}\right)\right) \\ &\geqslant \frac{\Delta T}{6}\exp\left(-\mathop{\mathbb{E}}_{\pi,\boldsymbol{\nu}}[N_i(T)]D_{\mathrm{KL}}(\nu_i||\nu'_i)\right) \end{aligned}$$

We recall that by the definition of i, we have that $\mathbb{E}_{\pi,\boldsymbol{\nu}}[N_i(T)] \leqslant \frac{T}{k-1}$. We now compute the Kullback-Leibler divergence between the two instances:

$$\begin{aligned} D_{\mathrm{KL}}(\nu_i||\nu'_i) &= \Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}\ln\left(\frac{\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}{\frac{1}{3}\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}\right) + \underbrace{(1-\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}})\ln\left(\frac{1-\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}{1-\frac{1}{3}\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}\right)}_{\leqslant 0} \\ &\leqslant \Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}\ln 3. \end{aligned}$$

Plugging this result, we finally get:

$$R_T(\pi, \boldsymbol{\nu}) + R_T(\pi, \boldsymbol{\nu}') \geqslant \frac{\Delta T}{6} \exp\left(-\frac{T}{k-1}\Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \ln 3\right).$$

We conclude the proof by noting that $\max\{x, y\} > \frac{1}{2}(x+y)$ and setting $\Delta = \frac{1}{2}\left(\frac{k-1}{T} u^{\frac{1}{\epsilon}} \frac{1}{\ln 3}\right)^{\frac{\epsilon}{1+\epsilon}}$. Finally, we have:

$$\max\{R_T(\pi, \boldsymbol{\nu}), R_T(\pi, \boldsymbol{\nu}')\} \geqslant c_3 k^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}},$$

for some constant $c_3 > 0$ independent of T, u, ϵ and k. ■

D.2.2. Estimator

Lemma 28 ((ϵ, u)-free Upper Confidence Bound)**.** *Let $\delta \in (0, 1/2)$ and $\mathbf{X} = \{X_1, \dots, X_s\}$ be a set of $s \in \mathbb{N}_{\geqslant 2}$ i.i.d. random variables satisfying $X_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, $\mu := \mathbb{E}[X_1]$, and $M > 0$ be a (possibly random) trimming threshold independent of $\mathbf{X}$. Then, under Assumption 6, it holds that:*

$$\mathbb{P}\left(\mu - \widehat{\mu}_s(\mathbf{X}; M) \leqslant \sqrt{\frac{2V_s(\mathbf{X}; M) \ln \delta^{-1}}{s}} + \frac{10M \ln \delta^{-1}}{s}\right) \geqslant 1 - 2\delta, \tag{4.10}$$

where $V_s(\mathbf{X}; M)$ is the sample variance of the trimmed random variables, defined as:

$$V_s(\mathbf{X}; M) := \frac{1}{s-1} \sum_{j \in [s]} (X_j \mathbb{1}_{\{|X_j| \leqslant M\}} - \widehat{\mu}_s(\mathbf{X}; M))^2. \tag{4.11}$$

Proof Since M is computed independently of $\mathbf{X}$, the trimmed samples $X_i \mathbb{1}_{\{|X_i| \leqslant M\}}$ remain

independent. Thus, with probability at least $1-\delta$, we have:

$$\begin{aligned}
\mu - \widehat{\mu}_s(\mathbf{X}; M) &= \mathbb{E}[X_1] - \frac{1}{s}\sum_{t=1}^{s} X_t \mathbb{1}_{|X_t|\leqslant M} \\
&= \frac{1}{s}\sum_{t=1}^{s}\left(\mathbb{E}[X_1] - \mathbb{E}\left[X_t \mathbb{1}_{|X_t|\leqslant M}\right]\right) + \frac{1}{s}\sum_{t=1}^{s}\left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t|\leqslant M}\right] - X_t \mathbb{1}_{|X_t|\leqslant M}\right) \\
&= \frac{1}{s}\sum_{t=1}^{s}\mathbb{E}[X_t \mathbb{1}_{|X_t|> M}] + \frac{1}{s}\sum_{t=1}^{s}\left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t|\leqslant M}\right] - X_t \mathbb{1}_{|X_t|\leqslant M}\right) \\
&\overset{(*)}{\leqslant} \frac{1}{s}\sum_{t=1}^{s}\left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t|\leqslant M}\right] - X_t \mathbb{1}_{|X_t|\leqslant M}\right) \\
&\overset{(**)}{\leqslant} \sqrt{\frac{2V_s(\mathbf{Y})\ln 2\delta^{-1}}{s}} + \frac{14M\ln 2\delta^{-1}}{3(s-1)} \\
&\leqslant \sqrt{\frac{2V_s(\mathbf{Y})\ln 2\delta^{-1}}{s}} + \frac{10M\ln 2\delta^{-1}}{s}
\end{aligned}$$

Note that in step $(*)$ we used Assumption 6 to make the first term vanish. In step $(**)$, instead, we used *empirical Bernstein inequality* (Maurer and Pontil, 2009) recalling that the trimmed random variables range in $[-M, M]$. We also use $\frac{1}{s-1} \leqslant \frac{2}{s}$ in the last step for $s \geqslant 2$. ■

Proposition 59 (Uniqueness of Solution of Equation (4.12), Wang et al. (2021)). *Let* $\mathbf{X} = \{X_1, \dots, X_s\}$ *be a set of real numbers. If:*

$$0 < c\ln\delta^{-1} < \sum_{j\in[s]} \mathbb{1}_{\{X_i\neq 0\}}, \tag{D.14}$$

then Equation (4.12) *admits a unique positive solution.*

Theorem 29 (Bounds on $\widehat{M}_s(\delta)$). *Let* $\delta \in (0, 1/2)$ *and* $\mathbf{X}' = \{X'_1, \dots, X'_s\}$ *be a set of* $s \in \mathbb{N}_{\geqslant 1}$ *i.i.d. random variables satisfying* $X'_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$*, and let* $\widehat{M}_s(\delta)$ *be the (random) positive root of Equation* (4.12) *with* $c > 2$*. Then, if* $\widehat{M}_s(\delta)$ *exists, with probability at least* $1-2\delta$*, it holds that:*

$$\widehat{M}_s(\delta) \leqslant \left(\frac{us}{(\sqrt{c}-\sqrt{2})^2\ln\delta^{-1}}\right)^{\frac{1}{1+\epsilon}} \quad \textit{and} \quad \mathbb{P}\left(|X_1| > \widehat{M}_s(\delta)\right) \leqslant (\sqrt{c}+\sqrt{2})^2\frac{\ln\delta^{-1}}{s}. \tag{4.15}$$

Proof The proof makes use of the concentration inequality for self-bounding random variables (Maurer, 2006; Maurer and Pontil, 2009). Let $M > 0$, for every $i \in [\![s]\!]$, we define the random

variable:

$$U_{i,M} \coloneqq \min\left\{\left(\frac{X_i}{M}\right)^2, 1\right\},$$

that ranges in $[0,1]$. Furthermore, let: $Z_M(\mathbf{X}) \coloneqq \sum_{i=1}^{s} U_{i,M}$, ranging in $[0,s]$. Let us denote $\overline{U}_M(\mathbf{X}) \coloneqq Z_M(\mathbf{X})/s$, we observe that, given these definitions, the equation we want to solve for non-zero roots becomes:

$$\overline{U}_M(\mathbf{X}) - \frac{c\ln\delta^{-1}}{s} = 0. \tag{D.15}$$

We start by showing that $Z_M(\mathbf{X})$ satisfies the assumptions of Theorem 13 of Maurer (2006), in particular, let $a \geqslant 1$, we have:

$$Z_M(\mathbf{X}) - \inf_{y\in\mathbb{R}} Z_M(\mathbf{X}_{y,k}) \leqslant 1, \quad \forall k \in [s], \tag{D.16}$$

$$\sum_{k=1}^{s}\left(Z_M(\mathbf{X}) - \inf_{y\in\mathbb{R}} Z_M(\mathbf{X}_{y,k})\right)^2 \leqslant aZ_M(\mathbf{X}), \tag{D.17}$$

where $\mathbf{X}_{y,k}$ is obtained by replacing with y the k-th element X_k of the set $\mathbf{X}$. Indeed, Equation (D.16) follows as:

$$Z_M(\mathbf{X}) - \inf_{y\in\mathbb{R}} Z_M(\mathbf{X}_{y,k}) = U_{k,M} - \inf_{y\in\mathbb{R}} \min\left\{\left(\frac{y}{M}\right)^2, 1\right\} = U_{k,M} \leqslant 1, \quad \forall k \in [s].$$

Similarly, we set $a = 1$ and obtain Equation (D.17) as follows:

$$\begin{aligned}
\sum_{k=1}^{s}\left(Z_M(\mathbf{X}) - \inf_{y\in\mathbb{R}} Z_M(\mathbf{X}_{y,k})\right)^2 &= \sum_{k=1}^{s}\left(U_{k,M} - \inf_{y\in\mathbb{R}} \min\left\{\left(\frac{y}{M}\right)^2, 1\right\}\right)^2 \\
&\leqslant \sum_{k=1}^{s} U_{k,M}^2 \\
&\leqslant \sum_{k=1}^{s} U_{k,M} \\
&= Z_M(\mathbf{X}),
\end{aligned}$$

since $U_{k,M} \leqslant 1$. Using Theorem 13 from Maurer (2006) with $a = 1$, for the right tail of the distribution, we have for every $\epsilon > 0$:

$$\mathbb{P}\left(\mathbb{E}[Z_M(\mathbf{X})] - Z_M(\mathbf{X}) > s\epsilon\right) \leqslant \exp\left(\frac{-\epsilon^2 s^2}{2\mathbb{E}[Z_M(\mathbf{X})]}\right)$$

By letting $\epsilon = \sqrt{\frac{2\mathbb{E}[\overline{U}_M(\mathbf{X})]\ln 2\delta^{-1}}{s}}$ and recalling the definition of $\overline{U}_M(\mathbf{X})$, we obtain:

$$\mathbb{P}\left(\mathbb{E}[\overline{U}_M(\mathbf{X})] - \overline{U}_M(\mathbf{X}) > \sqrt{\frac{2\mathbb{E}[\overline{U}_M(\mathbf{X})]\ln \delta^{-1}}{s}}\right) \leqslant \delta,$$

which implies, after some algebraic manipulations (see Theorem 10 of (Maurer and Pontil, 2009)), the following:

$$\mathbb{P}\left(\sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} - \sqrt{\overline{U}_M(\mathbf{X})} > \sqrt{\frac{2\ln \delta^{-1}}{s}}\right) \leqslant \delta.$$

A similar inequality holds for the left tail:

$$\mathbb{P}\left(Z_M(\mathbf{X}) - \mathbb{E}[Z_M(\mathbf{X})] > s\epsilon\right) \leqslant \exp\left(\frac{-\epsilon^2 s^2}{2\mathbb{E}[Z_M(\mathbf{X})] + \epsilon s}\right),$$

with similar steps, we obtain:

$$\mathbb{P}\left(\sqrt{\overline{U}_M(\mathbf{X})} - \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} > \sqrt{\frac{2\ln \delta^{-1}}{s}}\right) \leqslant \delta.$$

With a union bound over the two inequalities on the left and the right tail, we finally get:

$$\mathbb{P}\left(\left|\sqrt{\overline{U}_M(\mathbf{X})} - \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]}\right| > \sqrt{\frac{2\ln \delta^{-1}}{s}}\right) \leqslant 2\delta. \tag{D.18}$$

Let us now define $\widehat{M}_s(\delta)$ random variable corresponding to the solution of the equation:

$$\overline{U}_{\widehat{M}_s(\delta)}(\mathbf{X}) = \frac{c\ln \delta^{-1}}{s},$$

where $c > 0$. To control the bounds on $\widehat{M}$, we define the following auxiliary (non-random) quantities:

$$\sqrt{\overline{U}_M^+} := \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} + \sqrt{\frac{2\ln \delta^{-1}}{s}} \quad \text{and} \quad \sqrt{\overline{U}_M^-} := \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} - \sqrt{\frac{2\ln \delta^{-1}}{s}}. \tag{D.19}$$

Thanks to Equation D.18, we have, for every $M \geqslant 0$, that $\mathbb{P}(\overline{U}_M^- \leqslant \overline{U}_M(\mathbf{X}) \leqslant \overline{U}_M^+) \geqslant 1 - 2\delta$.

Furthermore, let $M^+(\delta), M^-(\delta) > 0$, the solutions of the following (non-random) equations:

$$U^+_{M^+(\delta)} = \frac{c \ln \delta^{-1}}{s} \qquad \text{and} \qquad U^-_{M^-(\delta)} = \frac{c \ln \delta^{-1}}{s}. \tag{D.20}$$

Since $\mathbb{P}(\overline{U}^-_M \leqslant \overline{U}_M(\mathbf{X}) \leqslant \overline{U}^+_M) \geqslant 1 - 2\delta$, it follows that $\mathbb{P}(M^-(\delta) \leqslant \widehat{M}_s(\delta) \leqslant M^+(\delta)) \geqslant 1 - 2\delta$. We now proceed at lower bounding $M^-(\delta)$ and upper bounding $M^+(\delta)$:

$$\begin{aligned}
\sqrt{\frac{c \ln \delta^{-1}}{s}} &= \sqrt{U^-_{M^-(\delta)}} && \text{(D.21)}\\
&= \sqrt{\mathbb{E}[\overline{U}_{M^-(\delta)}(\mathbf{X})]} - \sqrt{\frac{2 \ln \delta^{-1}}{s}} && \text{(D.22)}\\
&\geqslant \sqrt{\mathbb{P}\left(|X_1| \geqslant M^-(\delta)\right)} - \sqrt{\frac{2 \ln \delta^{-1}}{s}} && \text{(D.23)}\\
&\geqslant \sqrt{\mathbb{P}\left(|X_1| \geqslant \widehat{M}_s(\delta)\right)} - \sqrt{\frac{2 \ln \delta^{-1}}{s}}, && \text{(D.24)}
\end{aligned}$$

where the last but one inequality follows from:

$$\mathbb{E}[\overline{U}_M(\mathbf{X})] = \mathbb{E}\left[\min\left\{\left(\frac{X_1}{M}\right)^2, 1\right\}\right] \geqslant \mathbb{P}\left(\left(\frac{X_1}{M}\right)^2 \geqslant 1\right) = \mathbb{P}\left(|X_1| \geqslant M\right), \tag{D.25}$$

and the last inequality holds with probability $1-\delta$ and follows from the fact that $\widehat{M}_s(\delta) \geqslant M^-(\delta)$. Similarly, we have:

$$\begin{aligned}
\sqrt{\frac{c \ln \delta^{-1}}{s}} &= \sqrt{U^+_{M^+(\delta)}} && \text{(D.26)}\\
&= \sqrt{\mathbb{E}[\overline{U}_{M^-(\delta)}(\mathbf{X})]} + \sqrt{\frac{2 \ln \delta^{-1}}{s}} && \text{(D.27)}\\
&\leqslant \sqrt{\frac{u}{(M^+(\delta))^{1+\epsilon}}} + \sqrt{\frac{2 \ln \delta^{-1}}{s}} && \text{(D.28)}\\
&\leqslant \sqrt{\frac{u}{(\widehat{M}_s(\delta))^{1+\epsilon}}} + \sqrt{\frac{2 \ln \delta^{-1}}{s}}, && \text{(D.29)}
\end{aligned}$$

where the last but one inequality follows from:

$$\mathbb{E}[\overline{U}_M(\mathbf{X})] = \mathbb{E}\left[\min\left\{\left(\frac{X_1}{M}\right)^2, 1\right\}\right] \leqslant M^{-1-\epsilon}\mathbb{E}\left[|X_1|^{1+\epsilon}\right] \leqslant M^{-1-\epsilon}u, \tag{D.30}$$

and the last inequality holds with probability $1-\delta$ and follows from the fact that $\widehat{M}_s(\delta) \leqslant M^+(\delta)$.

Thus, with probability $1-2\delta$, we have for $c>2$:

$$\mathbb{P}\left(|X_1|>\widehat{M}_s(\delta)\right)\leqslant(\sqrt{c}+\sqrt{2})^2\frac{\ln\delta^{-1}}{s}\quad\text{and}\quad\widehat{M}_s(\delta)\leqslant\left(\frac{us}{(\sqrt{c}-\sqrt{2})^2\ln\delta^{-1}}\right)^{\frac{1}{1+\epsilon}}. \tag{D.31}$$

■

Theorem 30 ((ϵ,u)-dependent Concentration Bound)**.** *Let $\delta\in(0,1/4)$, $\mathbf{X}=\{X_1,\dots,X_{s/2}\}$, and $\mathbf{X}'=\{X'_1,\dots,X'_{s/2}\}$ be two independent sets of $s/2\in\mathbb{N}_{\geqslant 2}$ i.i.d. random variables satisfying $X_1\sim\nu\in\mathcal{P}_{HT}(\epsilon,u)$, $\mu:=\mathbb{E}[X_1]$, and let $\widehat{M}_s(\delta)$ be the (random) positive root of Equation* (4.12) *with $c=(1+\sqrt{2})^2$. Then, if $\widehat{M}_s(\delta)$ exists, it holds that:*

$$\mathbb{P}\left(\left|\widehat{\mu}_s(\mathbf{X};\widehat{M}_s(\delta))-\mu\right|\leqslant 8u^{\frac{1}{1+\epsilon}}\left(\frac{\ln\delta^{-1}}{s}\right)^{\frac{\epsilon}{1+\epsilon}}\right)\geqslant 1-4\delta. \tag{4.17}$$

Proof The result is obtained by combining an application of Bernstein's inequality and the bounds on the threshold $\widehat{M}_s(\delta)$ of Lemma 29. Furthermore since $\widehat{M}_s(\delta)$ is independent of $\mathbf{X}$, we can condition on the value of $\widehat{M}_s(\delta)$. With probability $1-\delta$, we have:

$$\begin{aligned}
\widehat{\mu}_s(\mathbf{X};\widehat{M}_s(\delta))-\mu&=\frac{1}{s}\sum_{i=1}^{s}X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}-\mathbb{E}[X_1]\\
&=\frac{1}{s}\sum_{i=1}^{s}\left(X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}-\mathbb{E}\left[X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}\right]\right)-\frac{1}{s}\sum_{i=1}^{s}\left(\mathbb{E}[X_1]-\mathbb{E}\left[X_t\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}\right]\right)\\
&=\frac{1}{s}\sum_{i=1}^{s}\left(X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}-\mathbb{E}\left[X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}\right]\right)-\frac{1}{s}\sum_{i=1}^{s}\mathbb{E}[X_i\mathbb{1}_{|X_i|>\widehat{M}_s(\delta)}]\\
&\leqslant\frac{1}{s}\sum_{i=1}^{s}\left(X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}-\mathbb{E}\left[X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}\right]\right)+\frac{1}{s}\sum_{i=1}^{s}\mathbb{E}[|X_i|\mathbb{1}_{|X_i|>\widehat{M}_s(\delta)}]\\
&\overset{(*)}{\leqslant}\frac{1}{s}\sum_{i=1}^{s}\left(X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}-\mathbb{E}\left[X_i\mathbb{1}_{|X_i|\leqslant\widehat{M}_s(\delta)}\right]\right)+\\
&\quad+\frac{1}{s}\sum_{i=1}^{s}\left(\mathbb{E}\left[|X_i|^{1+\epsilon}\right]^{\frac{1}{1+\epsilon}}\right)\left(\mathbb{E}\left[\left(\mathbb{1}_{|X_i|>\widehat{M}_s(\delta)}\right)^{\frac{1+\epsilon}{\epsilon}}\right]^{\frac{\epsilon}{1+\epsilon}}\right)\\
&\overset{(**)}{\leqslant}\sqrt{\frac{2\widehat{M}_s(\delta)^{1-\epsilon}u\ln\left(\delta^{-1}\right)}{s}}+\frac{\widehat{M}_s(\delta)\ln\left(\delta^{-1}\right)}{3s}+\frac{1}{s}\sum_{i=1}^{s}\left(u^{\frac{1}{1+\varepsilon}}\right)\left(\mathbb{E}\left[\mathbb{1}_{|X_i|>\widehat{M}_s(\delta)}\right]^{\frac{\epsilon}{1+\varepsilon}}\right)\\
&\leqslant\sqrt{\frac{2\widehat{M}_s(\delta)^{1-\epsilon}u\ln\left(\delta^{-1}\right)}{s}}+\frac{\widehat{M}_s(\delta)\ln\left(\delta^{-1}\right)}{3s}+u^{\frac{1}{1+\varepsilon}}\mathbb{P}\left(|X_i|>\widehat{M}_s(\delta)\right)^{\frac{\epsilon}{1+\varepsilon}},
\end{aligned}$$

where step $(*)$ follows from Hölder inequality, while step $(**)$ is a consequence of Bernstein's inequality for bounded random variables. To proceed further, we use Lemma 29 in union bound with the previously applied inequality. Thus, with probability at least $1-3\delta$, we have:

$$\begin{aligned}
&\widehat{\mu}_s(\mathbf{X};\widehat{M}_s(\delta))-\mu \leqslant \\
&\leqslant \sqrt{\frac{2\left(\frac{us}{(\sqrt{c}-\sqrt{2})^2\ln\delta^{-1}}\right)^{\frac{1-\epsilon}{1+\epsilon}} u\ln\left(\delta^{-1}\right)}{s}} + \frac{\left(\frac{us}{(\sqrt{c}-\sqrt{2})^2\ln\delta^{-1}}\right)^{\frac{1}{1+\epsilon}}\ln\left(\delta^{-1}\right)}{3s} \\
&\qquad + u^{\frac{1}{1+\varepsilon}}\left((\sqrt{c}+\sqrt{2})^2\frac{\ln\delta^{-1}}{s}\right)^{\frac{\epsilon}{1+\epsilon}} \\
&\leqslant \left(\frac{\sqrt{2}}{(\sqrt{c}-\sqrt{2})^{\frac{1-\epsilon}{1+\epsilon}}} + \frac{1}{3(\sqrt{c}-\sqrt{2})^{\frac{2}{1+\epsilon}}} + (\sqrt{c}+\sqrt{2})^{\frac{2\epsilon}{1+\epsilon}}\right) u^{\frac{1}{1+\epsilon}}\left(\frac{\ln\delta^{-1}}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \\
&\leqslant 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\ln\delta^{-1}}{s}\right)^{\frac{\epsilon}{1+\epsilon}},
\end{aligned}$$

where in the last passage we set $c=(1+\sqrt{2})^2$ and bounded the resulting expression for $\epsilon\in(0,1]$. A symmetric derivation leads to the second inequality. A union bound combined with renaming $s\leftarrow s/2$ and using $5.6\sqrt{2}\leqslant 8$, concludes the proof. ■

D.2.3. Upper Bound

Theorem 31 (Instance-Dependent Regret bound of `AdaR-UCB`)**.** *Let $\boldsymbol{\nu}\in\mathcal{P}_{HT}(\epsilon,u)^k$ and $T\in\mathbb{N}_{\geqslant 2}$ be the learning horizon. Under Assumption 6,* `AdaR-UCB` *suffers a regret bounded as:*

$$R_{\boldsymbol{\nu},T}(\pi^{\texttt{AdaR-UCB}})\leqslant\sum_{i:\Delta_i>0}\left[\left(120\left(\frac{u}{\Delta_i}\right)^{\frac{1}{\epsilon}}+\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\right)\ln\frac{T}{2}+20\Delta_i\right]. \tag{4.18}$$

Proof For notational convenience, in this derivation, we will perform the substitution $T\leftarrow\lfloor T/2\rfloor$ and $t\leftarrow\tau$. For every arm $i\in[k]$ and round $t\in[T]$, let us define the event:

$$\mathcal{E}_{i,t}:=\left\{\sum_{X\in\mathbf{X}'_i(t-1)}\mathbb{1}_{\{X\neq 0\}}-4\ln t^3>0\right\}. \tag{D.32}$$

Under event $\mathcal{E}_{i,t}$ we do not incur in the forced exploration (FE) in line 4 ensuring that every arm has collected at least $4\ln t^3$ nonzero samples in $\mathbf{X}'_i$. Thus, we can decompose the expected

number of pulls as follows:

$$\mathbb{E}[N_i^{\text{ALL}}(T)] = \mathbb{E}\left[\sum_{t\in[T]} \mathbb{1}_{\{I_t=i \text{ and } \mathcal{E}_{i,t}\}}\right] + \mathbb{E}\left[\sum_{t\in[T]} \mathbb{1}_{\{I_t=i \text{ and } \mathcal{E}_{i,t}^{\complement}\}}\right] \tag{D.33}$$

$$= \mathbb{E}[N_i(T)] + \mathbb{E}[N_i^{\text{FE}}(T)]. \tag{D.34}$$

Part I: Bounding the expected number of pulls for forced exploration. We first bound the expected number of pulls $\mathbb{E}[N_i^{\text{FE}}(T)]$ due to the forced exploration. Considering only the samples collected due to forced exploration, thanks to independence among these samples, we can see the required number of pulls as a sum of geometric random variables. Thus, we can compute an upper bound on the expectation as:

$$\mathbb{E}_{\nu_i}[N_i^{\text{FE}}(T)] \leqslant \frac{4\ln T^3}{\mathbb{P}_{\nu_i}(|X|>0)}. \tag{D.35}$$

Part II: Bounding the expected number of pulls for optimistic exploration. We define for every arm $i \in [k]$ and every round $t \in [T]$, the upper confidence bound as:

$$B_i(t) = \widehat{\mu}_i(t) + \sqrt{\frac{2V_i(t)\ln t^3}{N_i(t-1)}} + \frac{10\widehat{M}_i(t)\ln t^3}{N_i(t-1)},$$

where $N_i(t-1)$ is the number of times arm i has been pulled up to time $t-1$, i.e., $N_i(t-1) = |\mathbf{X}_i(t-1)|$. We now show that if $I_t = i$, for an arm i such that $\Delta_i > 0$, then, one of the following four inequalities is true:

$$\text{either } B_1(t) \leqslant \mu^*, \tag{D.36}$$

$$\text{or} \quad \widehat{\mu}_i(t) > \mu_i + 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}}, \tag{D.37}$$

$$\text{or} \quad N_i(t-1) < 20\left(\frac{u}{\Delta_i^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}\ln t^3, \tag{D.38}$$

$$\text{or } \sqrt{V_i(t)} > \sqrt{\mathbb{E}[V_i(t)]} + 2\widehat{M}_i(t)\sqrt{\frac{\ln t^3}{N_i(t-1)}}, \tag{D.39}$$

$$\text{or } \widehat{M}_i(t) \geqslant \left(\frac{uN_i(t-1)}{\ln t^3}\right)^{\frac{1}{1+\epsilon}}. \tag{D.40}$$

Indeed, assume that all five inequalities are false. Then we have

$$\begin{aligned} B_1(t) &\overset{(D.36)}{>} \mu^* = \mu_i + \Delta_i \\ &\overset{(D.37)}{\geqslant} \widehat{\mu}_i(t) - 5.6u^{\frac{1}{1+\epsilon}} \left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \Delta_i \\ &\overset{(*)}{\geqslant} \widehat{\mu}_i(t) + \sqrt{\frac{2V_i(t)\ln t^3}{N_i(t-1)}} + \frac{10\widehat{M}_i(t)\ln t^3}{N_i(t-1)} \\ &= B_i(t). \end{aligned}$$

The step marked with $(*)$ is a consequence of the fact that both (D.38), (D.39) and (D.40) are false. In particular, we need to show that

$$\Delta_i \geqslant 5.6u^{\frac{1}{1+\epsilon}} \left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2V_i(t)\ln t^3}{N_i(t-1)}} + \frac{10\widehat{M}_i(t)\ln t^3}{N_i(t-1)}. \tag{$*$}$$

To do so, we make use of the following inequality derived by exploiting the independence between $\mathbf{X}_i(t-1)$ and $\mathbf{X}'_i(t-1)$:

$$\mathbb{E}[V_i(t)] \leqslant \mathbb{E}\left[X^2 \mathbb{1}_{|X|\leqslant \widehat{M}_i(t)}\right] \leqslant \mathbb{E}\left[|X|^{1+\epsilon}\right] \widehat{M}_i(t)^{1-\varepsilon} \leqslant u\widehat{M}_i(t)^{1-\epsilon}. \tag{D.41}$$

Now, we make use of the fact that (D.38), (D.39), and (D.40) are false together with (D.41):

$$
\begin{aligned}
\Delta_i &\overset{\text{(D.38)}}{\geqslant} 20u^{\frac{1}{1+\epsilon}}\left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} \\
&\geqslant (5.6+\sqrt{2}+10+2\sqrt{2})u^{\frac{1}{1+\epsilon}}\left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} \\
&= 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2\ln t^3 u\left(\frac{uN_i(t-1)}{\ln t^3}\right)^{\frac{1-\epsilon}{1+\epsilon}}}{N_i(t-1)}} + \frac{(10+2\sqrt{2})\left(\frac{uN_i(t-1)}{\ln t^3}\right)^{\frac{1}{1+\epsilon}}\ln t^3}{N_i(t-1)} \\
&\overset{\text{(D.40)}}{\geqslant} 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2\ln t^3 u\widehat{M}_i(t)^{1-\epsilon}}{N_i(t-1)}} + \frac{(10+2\sqrt{2})\widehat{M}_i(t)\ln t^3}{N_i(t-1)} \\
&\overset{\text{(D.41)}}{\geqslant} 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2\mathbb{E}[V_i(t)]\ln t^3}{N_i(t-1)}} + \frac{(10+2\sqrt{2})\widehat{M}_i(t)\ln t^3}{N_i(t-1)} \\
&\overset{\text{(D.39)}}{\geqslant} 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\ln t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2\ln t^3}{N_i(t-1)}}\left[\sqrt{V_i(t)} - 2\widehat{M}_i(t)\sqrt{\frac{\ln t^3}{N_i(t-1)}}\right] \\
&\qquad + \frac{(10+2\sqrt{2})\widehat{M}_i(t)\ln t^3}{N_i(t-1)} \\
&\geqslant 5.6u^{\frac{1}{1+\epsilon}}\left[\frac{\ln t^3}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{V_i(t)\ln t^3}{N_i(t-1)}} + \frac{10\widehat{M}_i(t)\ln t^3}{N_i(t-1)}.
\end{aligned}
$$

(∗)

Finally, as a consequence of $(*)$, we have $B_1(t) > B_i(t)$ but this is a contradiction since $T_t = i$. Thus, statements (D.36) to (D.40) cannot be false simultaneously. We now proceed with a union bound over all the possible values of $N_i(t-1)$ and of the previously introduced concentration inequalities to bound with $\frac{1}{t^3}$ the probabilities of events (D.36), (D.37), (D.39), and (D.40) to be true:

$$\begin{aligned}\mathbb{P}\left(\exists N_i(t-1) \in [t] \,:\, \{\text{(D.36) is true}\} \text{ or } \{\text{(D.37) is true}\} \text{ or } \{\text{(D.39) is true}\} \text{ or } \{\text{(D.40) is true}\}\right) \leqslant \\ \leqslant 6\sum_{s=1}^{t}\frac{1}{t^3} = \frac{6}{t^2},\end{aligned}$$

where for (D.39), we used the second inequality of Theorem 10 of (Maurer and Pontil, 2009) (bounding $1/(n-1) \leqslant 2/n$) and for (D.40), we used Theorem 29. To proceed, we introduce the quantity:

$$v \coloneqq \left\lceil 60\left(\frac{u}{\Delta_i^{1+\varepsilon}}\right)^{\frac{1}{\varepsilon}} \ln T \right\rceil.$$

It's now time to bound the expected number of times each arm is pulled:

$$\begin{aligned}\mathbb{E}[N_i(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i \text{ and } \mathcal{E}_{i,t}\}}\right] \\ &\leqslant v + \mathbb{E}\left[\sum_{t=v+1}^{T} \mathbb{1}_{\{I_t = i \text{ and } \{\text{(D.38) is false}\}\}}\right] \\ &\leqslant v + \mathbb{E}\left[\sum_{t=v+1}^{T} \mathbb{1}_{\{I_t = i \text{ and } \{\text{(D.36) or (D.37) or (D.39) or (D.40) is true}\}\}}\right] \\ &\leqslant v + \sum_{t=v+1}^{T}\frac{6}{t^2} \\ &\leqslant v + 10.\end{aligned} \tag{D.42}$$

We now conclude the proof using the regret decomposition, considering the forced exploration through Equation (D.35) and that the effective number of pulls is doubled:

$$R_T(\texttt{AdaR-UCB}, \boldsymbol{\nu}) \leqslant \sum_{i:\Delta_i>0}\left[\left(120\left(\frac{u}{\Delta_i}\right)^{\frac{1}{\epsilon}} + \frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\right)\ln\frac{T}{2} + 20\Delta_i\right].$$

■

Theorem 32 (Worst-Case Regret bound of AdaR-UCB)**.** *Let $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^k$ and $T \in \mathbb{N}_{\geqslant 2}$ be the*

learning horizon. Under Assumption 6, `AdaR-UCB` *suffers a regret bounded as:*

$$R_{\boldsymbol{\nu},T}(\pi^{\texttt{AdaR-UCB}}) \leqslant 46\left(k\ln\frac{T}{2}\right)^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\ln\frac{T}{2} + 20\Delta_i\right).$$

Proof Let us fix $\Delta > 0$, to be chosen later. We have:

$$\begin{aligned}
R_T(\texttt{AdaR-UCB},\boldsymbol{\nu}) &= \sum_{i\in[k]}\Delta_i\left(2\mathbb{E}[N_i(T/2)] + \mathbb{E}_{\nu_i}[N_i^{\mathrm{FE}}(T/2)]\right)\\
&= \sum_{i:\Delta_i\leqslant\Delta} 2\Delta_i\mathbb{E}[N_i(T/2)] + \sum_{i:\Delta_i>\Delta} 2\Delta_i\mathbb{E}[N_i(T/2)] + \sum_{i:\Delta_i>0}\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\ln\frac{T}{2}\\
&\leqslant \Delta T + \sum_{i:\Delta_i>\Delta} 2\Delta_i\left(60\left(\frac{u}{\Delta_i^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}\ln\frac{T}{2} + 10\right) + \sum_{i:\Delta_i>0}\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\ln\frac{T}{2}\\
&\leqslant \Delta T + 2k\left(60\left(\frac{u}{\Delta}\right)^{\frac{1}{\epsilon}}\ln\frac{T}{2}\right) + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\ln\frac{T}{2} + 20\Delta_i\right)\\
&\overset{(*)}{\leqslant} 120^{\frac{\epsilon}{1+\epsilon}}(1+\epsilon)\epsilon^{-\frac{\epsilon}{1+\epsilon}}\left(k\ln\frac{T}{2}\right)^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\ln\frac{T}{2} + 20\Delta_i\right)\\
&\overset{(**)}{\leqslant} 46\left(k\ln\frac{T}{2}\right)^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\ln\frac{T}{2} + 20\Delta_i\right),
\end{aligned}$$

where the step marked with $(*)$ follows by a proper choice of Δ minimizing the bound:

$$T - 120ku^{\frac{1}{\epsilon}}\epsilon^{-1}\Delta^{-\frac{1+\epsilon}{\epsilon}}\ln\frac{T}{2} = 0 \implies \Delta = \left(\frac{120ku^{\frac{1}{\epsilon}}\ln\frac{T}{2}}{\epsilon T}\right)^{\frac{\epsilon}{1+\epsilon}},$$

and step marked with $(**)$ follows by bounding simple numerical bounds. ■

D.3. Efficient Numerical Resolution of Equation (4.12)

In this appendix, we present a computationally efficient strategy that can be implemented in Algorithm 11 to execute line 7, *i.e.*, the solution of the root-finding problem. In particular, to solve the equation:

$$f_s(\mathbf{X}';M,\delta) := \frac{1}{s}\sum_{j\in[s]}\frac{\min\{(X'_j)^2,M^2\}}{M^2} - \frac{c\ln\delta^{-1}}{s} = 0. \tag{4.12}$$

[ht!] Reward set $\mathbf{X}' = \{X'_1, \dots, X'_s\}$, time counter τ, machine tolerance $\eta > 0$.
Initialize counter $h \leftarrow 0$, initial guess $x_0 \leftarrow \eta$, initial value $y_0 \leftarrow f_s(\mathbf{X}'; x_0, \tau^{-3})$.
while $y_h > 0$ **do**
$x_{h+1} \leftarrow 2x_h$
$y_{h+1} \leftarrow f_s(\mathbf{X}'; x_{h+1}, \tau^{-3})$
$h \leftarrow h + 1$

end

Return x_h .

We propose Algorithm D.3 to find an upper bound $\bar{M}_s(\tau^{-3})$ on the true solution $\widehat{M}_s(\tau^{-3})$ which is based on *bisection*. The strategy works as follows. We provide the minimum numerical tolerance of our machine $\eta > 0$, start from an initial guess $x_0 = \eta$, then, if this guess is an underestimation (*i.e.*, $f_s(\cdot, x_0)$ yields a positive value y_0) we proceed to iteratively double our guess until the real threshold has been passed (lines D.3-D.3). In line D.3, we return the final guess x_h. If the initial guess is already an overestimation of the threshold (*i.e.*, $f_s(\cdot, x_0)$ yields a negative value y_0), we simply have $x_0 = x_h = \eta$.

We point out that, by construction, the output of Algorithm D.3 can be *at most* two times the true solution to Equation (4.12), *i.e.*, $\bar{M}_s(\tau^{-3}) \leqslant 2\widehat{M}_s(\tau^{-3})$. Thus, regret guarantees for Algorithm 11 remain the same (up to numerical constants) even when performing this approximation of the threshold. In particular, in the proof of Theorem 31, we can modify (D.40) as follows:

$$\bar{M}_i(t) \geqslant 2\left(\frac{uN_i(t-1)}{3}\right)^{\frac{1}{1+\epsilon}},$$

and the final result remains the same up to multiplicative constants.

We now characterize the computational complexity of Algorithm D.3, *i.e.*, the maximum number of steps to be performed before returning a solution.

Proposition 60 (Upper Bound on the Number of Steps of Algorithm D.3)**.** *Let η be the minimum numerical tolerance, and assume $\eta \leqslant \widehat{M}_s(\tau^{-3})$. Then, in at most $\bar{h}_{\eta,\tau}(\epsilon, u)$ steps such that:*

$$\bar{h}_{\eta,\tau}(\epsilon, u) = \log_2\left(\frac{1}{\eta}\left(\frac{us}{\log(\tau^3)}\right)^{\frac{1}{1+\epsilon}}\right),$$

Algorithm D.3, returns a solution $x_{\bar{h}_{\eta,\tau}(\epsilon,u)}$ *s.t.*

$$\mathbb{P}\left(\frac{x_{\bar{h}_{\eta,\tau}(\epsilon,u)}}{\widehat{M}_s(\tau^{-3})} \in [1,2]\right) \geqslant 1 - \frac{2}{\tau^3}.$$

Proposition 60 states an upper bound for the number of steps of Algorithm D.3 as a function of both ϵ and u. However, we remark that these two are not required as input to the numerical solver. Moreover, it emerges a dependence on the inverse of the numerical tolerance of the machine on which the algorithm is run. Thanks to the logarithm, this dependence hardly becomes an issue. If we consider a very small tolerance of 10^{-16} (which is the standard tolerance of many programming languages) the number of steps becomes:

$$\bar{h}_{\eta,\tau}(\epsilon,u) = \log_2\left(\left(\frac{us}{\log(\tau^3)}\right)^{\frac{1}{1+\epsilon}}\right) + 16\log_2(10),$$

which is totally reasonable.

E | Regret Minimization in Piecewise-Stationary Heavy-Tailed Bandits

E.1. Proofs

Theorem 4.5 (Regret Lower Bound for the HTPS Bandit Problem)**.** *For any fixed policy π, we have*

$$\sup_{\nu \in \mathcal{B}_{(v,\epsilon,\Upsilon)}} R_T(\pi) \geqslant \frac{1}{25}(k\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}. \tag{4.20}$$

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 $\boldsymbol{i}^*$:

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

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:

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

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

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

that implies

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

Combining Equation (E.1) and Equation (E.2), we get

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

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

Remark 14. *In the proof of Theorem 4.5, 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., *$T \geqslant 2\Upsilon\lceil L_j \frac{k}{\eta} \rceil$, such that Assumption 4.9 holds. This proves that our assumption does not make the problem easier from a regret minimization perspective.*

Proposition 4.1 (Detection Delay of `Catoni-FCS-detector`)**.** *Consider a CPD problem with observations $\{X_t\}_{t\in\mathbb{N}}$ drawn i.i.d. from $P_0 \in \mathcal{H}_{\epsilon,v} \cap \mathcal{P}^{\mu_0}$ for $t \leqslant t_c$ and from $P_1 \in \mathcal{H}_{\epsilon,v} \cap \mathcal{P}^{\mu_1}$ for $t > t_c$. Let $\delta := |\mu_1 - \mu_0|$. Suppose that there exists a known upper bound T of the change point ($t_c \leqslant T$). Let $n_{min} := 68\log(T^{\frac{1+\epsilon}{\epsilon}})$ and suppose $t_c \geqslant n_{min}$ large enough s.t. $w(t_c, P_0, \gamma) \leqslant \frac{\delta}{2}$. Set $\gamma = \frac{2}{T^3}$. Then, there exists a predictable sequence $\{\lambda_i\}_{i=1}^T$ s.t.* `Catoni-FCS-detector` *enjoys* ***(i)*** $\mathbb{P}_{t_c}\left((\tau - t_c)^+ \leqslant \mathcal{O}\left(v^{\frac{1}{\epsilon}} \frac{\log(T)}{\delta^{\frac{1+\epsilon}{\epsilon}}}\right)\right) \geqslant 1 - \frac{14}{T}$ *and* ***(ii)*** $\mathbb{P}_{t_c}(\tau < t_c) \leqslant \frac{14}{T}$.

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

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

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

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

and, subsequently,

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

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

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

which implies

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

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

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

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:

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

and using Markov's ineqality we get:

$$\forall m \in \mathbb{R}, \quad \mathbb{P}\left(f_t(m) \leqslant B_t^{+}(m)\right) \geqslant 1 - \frac{\theta}{2}$$
$$\forall m \in \mathbb{R}, \quad \mathbb{P}\left(f_t(m) \geqslant B_t^{-}(m)\right) \geqslant 1 - \frac{\theta}{2}.$$

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

$$B_t^+(\widetilde{m}_t) = -\sum_{i=1}^{t} \frac{v\lambda_i^{1+\epsilon}}{1+\epsilon} - \log\left(\frac{2}{\gamma}\right) = f_t(\max\{CI_t^\phi\}) \tag{E.3}$$

also satisfies

$$\mathbb{P}\left(f_t(\widetilde{m}_t) \leqslant f_t(\max\{CI_t^\phi\}\right) \geqslant 1-\frac{\theta}{2}, \tag{E.4}$$

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 (E.4) implies that:

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

Note that Equation (E.3) admits solutions if and only if

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

Finally, we conclude this step by bounding

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

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

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

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

$$\mathbb{P}\left(w(t,\mu,\gamma) \leqslant 2\frac{\sum_{i=1}^{t} 10v\lambda_i^{1+\epsilon} + 2(1+\epsilon)\log\left(\frac{2}{\gamma}\right) + 2(1+\epsilon)\log\left(\frac{2}{\theta}\right)}{\sum_{i=1}^{t}\lambda_i}\right) \geqslant 1-\theta. \tag{E.6}$$

Step 4 (Stitching) We now discuss the choice of the sequence $\{\lambda_t\}_{t\geqslant 1}$. 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 $t_j < t \leqslant t_{j+1}$, we set $\lambda_i = \Lambda_j$ for every $i \in [t]$. Assume

$\theta = \frac{\gamma}{4}$. For every $t_j < t \leqslant t_{j+1}$, we have

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

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

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

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

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

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 $\mathcal{W}_{1:t} \coloneqq \bigcap_{i=1}^{t}\mathcal{W}_i$. 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 $\mathbb{P}(\mathcal{W}_{1:t}) \geqslant 1 - \frac{t(t+1)\gamma}{8}$.

Finally, note that $\mathcal{W}_{a:b} \subset \mathcal{W}_{a':b'}$, for every $a' > a$ and $b' < b$. Thus $\mathbb{P}(\mathcal{W}_{a:b}) \geqslant 1 - \frac{T(T+1)\gamma}{8}$ 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 $\mathbb{P}(\mathcal{E}_t^T) \geqslant 1 - (T-t)\gamma$. From now on, we continue by setting $\gamma = \frac{2}{T^3}$.

Step 6 (Verifying condition (E.5)**)** For every $t \in [T]$, we use the previously defined values for $\{\lambda_i\}_{i=1}^t$, γ and θ, and solve inequality (E.5). We obtain that it is satisfied for every $t \geqslant 68 \log\left(T^{\frac{1+\epsilon}{\epsilon}}\right) = n_{min}$, 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 (E.5). 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

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

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

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

We first bound $2\log(\log(e^2 n)) \leqslant 3\log(\log(n))$, that holds under the trivial requirements that $\log(n) \geqslant 2$ and $T \geqslant 2$. 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:

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

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

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

and note that $n_0 \leqslant n_1 + n_2$. 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

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

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

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

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

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

As a consequence, we can rewrite:

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

which immediately implies that

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

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

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

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 $\mathbb{P}\left((\mathcal{E}_1^{t_c})^C\right) \leqslant \frac{1}{T}$. ■

Lemma 61. *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* $\mathbb{P}\left(\mathcal{G}_T^C\right) \leqslant \frac{15k\Upsilon}{T}$.

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 4.9, the length of every epoch is at least $2\left\lceil L_j \frac{k}{\eta} \right\rceil$.

Let $\mathcal{L}^{(j)} := \left\{\forall m \leqslant j : \tau_m \in \left\{t_c^{(m)}, \dots, t_c^{(m)} + \left\lceil L_m \frac{k}{\eta} \right\rceil\right\}\right\}$ 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 4.1, we have:

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

■

Theorem 4.10 (Regret Upper Bound of R-CPD-UCB)**.** *Under Assumption 4.9,* R-CPD-UCB

suffers an expected cumulative regret bounded as:

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

(4.24)

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:

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

where the second inequality follows from Lemma 61. We can now focus on bounding the first addendum. We decompose it as follows:

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

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:

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

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

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

which yields, by a recursive application:

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

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

Let π_s be the `Robust UCB` policy with median-of-means estimator (Bubeck et al., 2013a, Section 2.2). Under Assumption 4.9, `R-CPD-UCB` suffers an expected cumulative regret bounded as:

$$\mathbb{E}[R^{\pi^{R\text{-}CPD\text{-}UCB}}(T)] \leqslant \mathcal{O}\left((A) + \underbrace{\sum_{j=1}^{\Upsilon} \sum_{i:\Delta_i^{(j)}>0} \frac{v^{\frac{1}{\epsilon}} \log(|E_j|)}{(\Delta_i^{(j)})^{\frac{1}{\epsilon}}}}_{(B_1)\text{ Robust UCB Regret (Instance Dependent)}} + (C) \right). \tag{4.26}$$

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

$$R_T(\pi^{R\text{-}CPD\text{-}UCB}) \leqslant \widetilde{\mathcal{O}}((A) + \underbrace{(k\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}}_{(B_2)\text{ Robust UCB Regret (Instance Independent)}} + (C)). \tag{4.27}$$

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. (2013a)). Equation (4.27) necessitates an additional step using Jensen Inequality:

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

■

Let π_s be the `Robust UCB` policy with median-of-means estimator from (Bubeck et al., 2013a, 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 4.9, `R-CPD-UCB` using η_{j+1} after the j-th detection suffers an expected cumulative regret bounded as:

$$R_T(\pi^{R\text{-}CPD\text{-}UCB}) \leqslant \mathcal{O}\left(\frac{v^{\frac{1}{\epsilon}}\sqrt{k\Upsilon T\log(T)}}{\eta_0 \delta_{min}^{\frac{1+\epsilon}{\epsilon}}}\Delta_{max} + \frac{k\Upsilon v^{\frac{1}{\epsilon}}\log(T/\Upsilon)}{\Delta_{min}^{\frac{1}{\epsilon}}} \right). \tag{4.28}$$

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

we have:

$$R_T(\pi^{R\text{-}CPD\text{-}UCB}) \leqslant \widetilde{\mathcal{O}}(\ (k\Upsilon)^{\frac{\epsilon}{1+\epsilon}}(vT)^{\frac{1}{1+\epsilon}}). \tag{4.29}$$

Proof First, note that the proof of Theorem 4.10 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 (4.25) is a trivial consequence of the fact that $\eta_{\Upsilon+1} \geqslant \eta_j$ for every $j \leqslant \Upsilon$, due to the monotonocity of the sequence. Moreover, we bound $\Delta_{max}^{(j)} \leqslant \Delta_{max}$.

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

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

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

E.2. Additional Related Works on Non-Stationary MABs

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

E.2.1. 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 sub-gaussian. 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.

E.2.2. Bounded Variation and Monotonically Non-stationary MABs

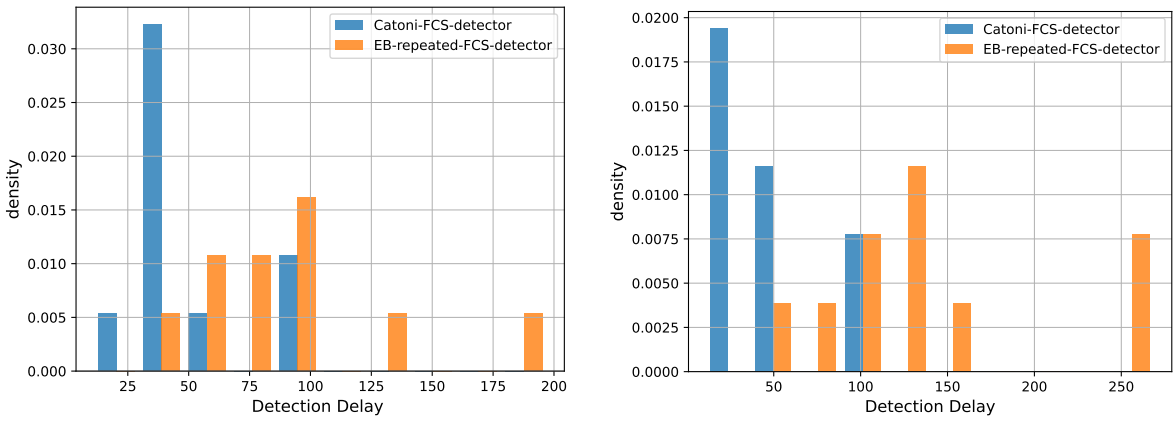
Another setting of interest is non-stationary MABs with *bounded variations*. In this setting, the rewards' distributions changes are less restricted, and the focus moves from the number of changes to the total amount of change V_T. In Besbes et al. (2014), the authors propose `Rexp3`, an algorithm that 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. (2024c). 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.

E.3. Additional Numerical Evaluations

In this appendix, we provide additional details on the experimental evaluation of Section 3.1.6 and additional experimental campaigns in synthetic environments.

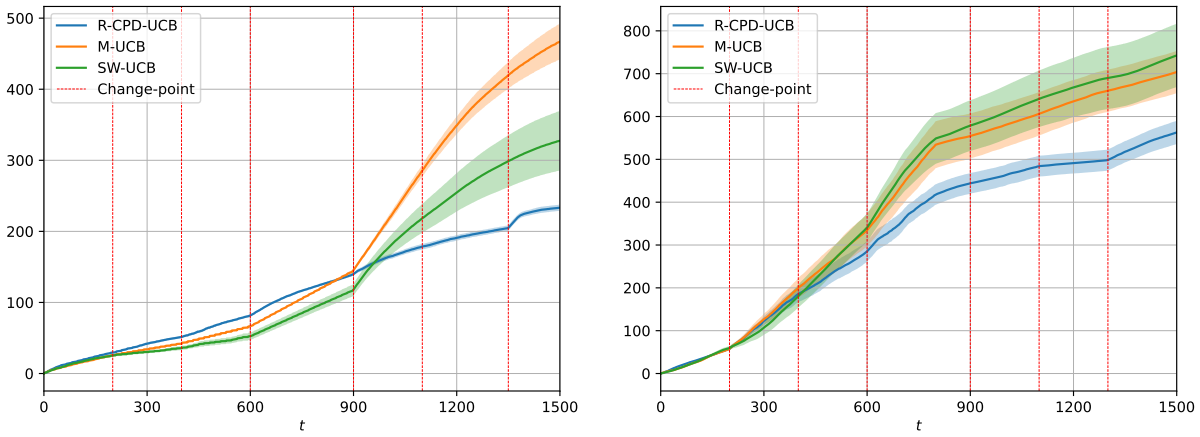
E.3.1. Detection Delay Analysis

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 (2023a), 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 E.1, we report the distribution of the detection delay of both algorithms 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.



(a) Gaussian samples. (b) Laplace samples.

Figure E.1: Distribution delay distribution over 20 trials.



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

Figure E.2: 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.

E.3.2. 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 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 4.9. We use $\sigma = 1$ and the means reported in Table E.3 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 E.4. The optimal actions change 4 times. In Figure E.6 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	$1, 7$

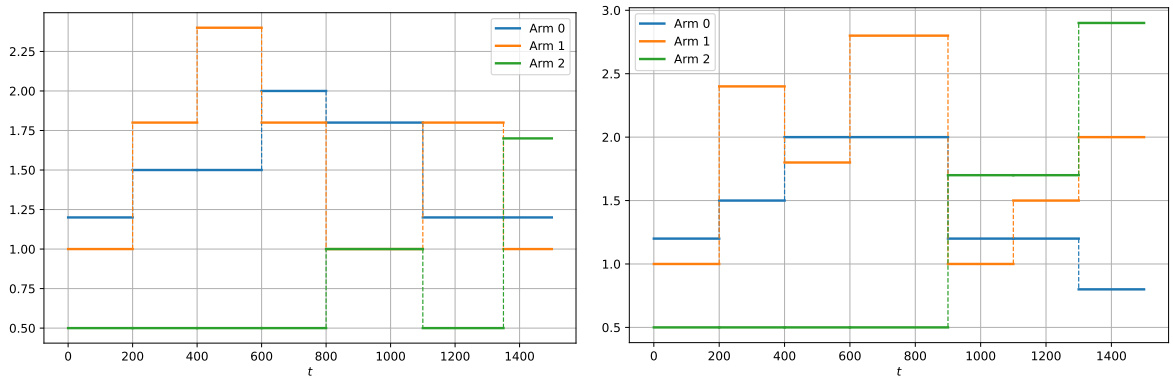
Figure E.3: Gaussian instance.

	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

Figure E.4: Pareto instance.

Figure E.5: Mean rewards per epoch. Cells highlighted in yellow contain the optimal reward for the corresponding epoch.

Results In Figure E.2, 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 E.2b). 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 4.9 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 4.9 is, in practice, is not very limiting.



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

Figure E.6: Average rewards per epoch.

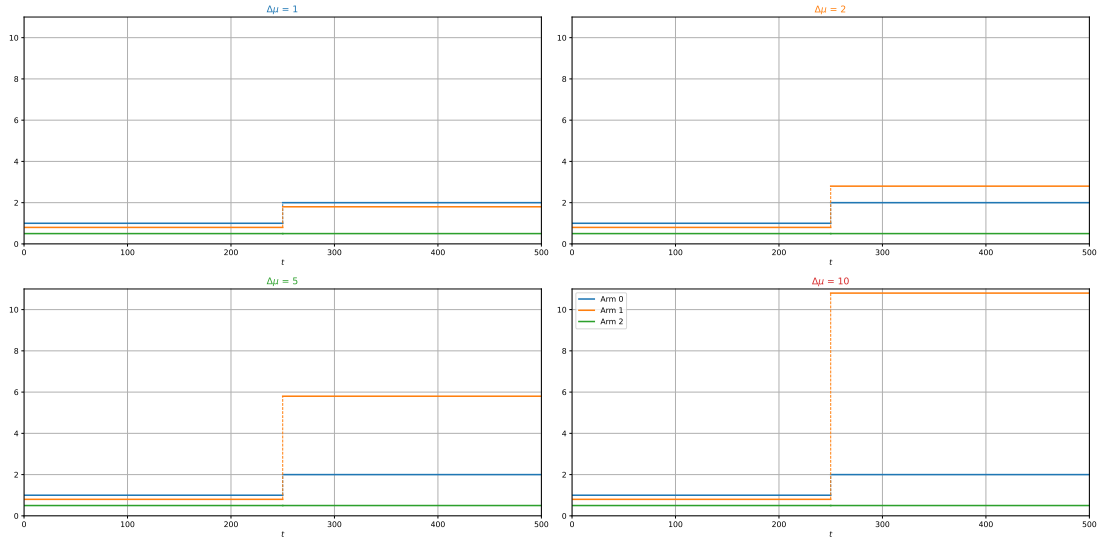
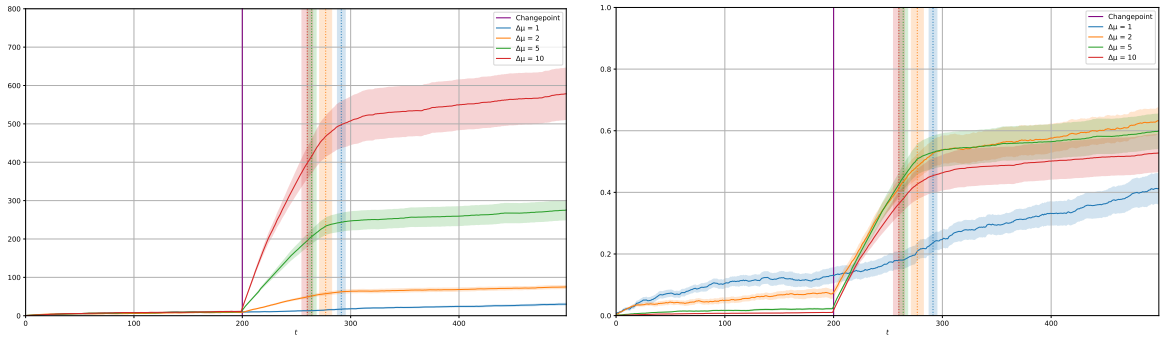


Figure E.7: HTPS MABs with different magnitudes of change. For all instances: $\Upsilon = 1$, $k = 3$, $\delta \in \{1, 2, 5, 10\}$. Pareto noise with $\epsilon < 1$ and $v = 3$.

E.3.3. 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 E.7, we report the means of every action in every epoch for the four HTPS MABs.



(a) Cumulative Regret. (b) Rescaled Cumulative Regret.

Figure E.8: 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.

Results In Figure E.8, 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 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.

E.3.4. 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. (2013a) 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

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

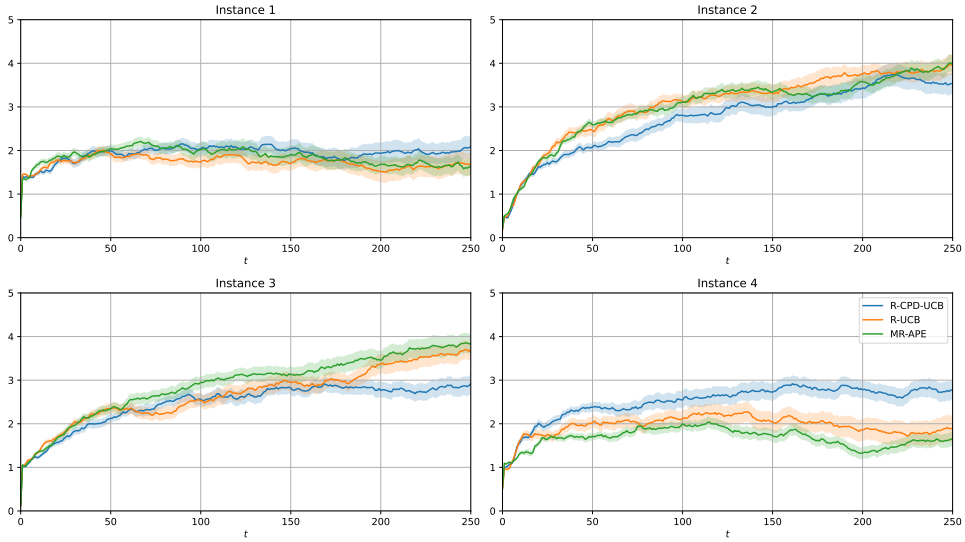


Figure E.9: Cumulative regrets of `R-CPD-UCB`, `Robust UCB` Bubeck et al. (2013a), and `MR-APE` Lee and Lim (2022) on the four HT MABs defined in Table 3. We performed 20 trials for each instance and reported mean $\pm$ std.

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 E.9, 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.

E.4. 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 E.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 (4.23) 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. (2022a), 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 $t' \leqslant T$ 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 $t, t' \leqslant T$. ■

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

List of Figures

List of Tables