POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Online Model Selection with Stochastic Rising Bandits

Tesi di Laurea Magistrale in
Computer Science Engineering - Ingegneria Informatica

Author: **Matteo Pirola**

Student ID: 940303
Advisor: Prof. Marcello Restelli
Co-advisors: Francesco Trovò, Alberto Maria Metelli
Academic Year: 2020-21

Abstract

This work lies in the field of stochastic Multi-Armed Bandits (MABs), i.e., the set of those online learning techniques which sequentially select an action (a.k.a. arm) observing only the feedback given by their choice (a.k.a. reward). A particular case arises when the arms' expected rewards (a.k.a. payoffs) are non-stationary and evolve as monotonic non-decreasing functions. We study that scenario in both the *restless* and the *rested* MAB formulation, meaning that the reward evolution is triggered by the natural flow of time or by the agent choices, respectively. This is the case of *online model selection* tasks, in which one would like to optimize the learning strategy during the learning phase itself. The assumptions under the *rising* bandit problem allow designing specific algorithms which exploit the regularity of the payoffs to provide tight regret guarantees. We design the `R-less-UCB` and `R-ed-UCB` algorithms, respectively for the rising restless and rising rested cases, providing a regret bound made of a problem-dependent component and a problem-independent one, which, under certain mild assumptions on the evolution of the reward function, is of order $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$, being T the learning horizon. Finally, we investigate the effectiveness of the proposed solutions by empirically comparing our novel approaches to state-of-the-art non-stationary bandit algorithms, using both synthetic and real-world data.

Keywords: Multi Armed Bandits, Online Model Selection, Restless Bandit, Rested Bandit, Rising Bandit

Abstract in lingua italiana

Questa tesi si pone nell'ambito dei Multi-Armed-Bandits (MABs) stocastici, ovvero l'insieme di quelle tecniche di selezione *online* che riescono a scegliere un'azione (in gergo tecnico chiamata *arm*) osservando solamente il risultato (*ricompensa*) delle loro scelte precedenti. Uno scenario particolare è quello in cui il valore atteso del guadagno (*payoff*) è non-stazionario ed evolve come una funzione monotona non-decrescente. Studiamo questo scenario sia nella formulazione di *restless* MAB sia in quella di *rested* MAB, ovvero quando l'evoluzione del guadagno è causata rispettivamente dallo scorrere del tempo o dalle scelte dell'agente. È proprio quest'ultimo il caso dei problemi di *selezione dei modelli online*, in cui si vuole ottimizzare il modello di learning durante la fase stessa di learning. Le assunzioni alla base di questa formulazione di *rising* bandits permettono di creare algoritmi specifici che, sfruttando le regolarità dei payoff, forniscono garanzie sul regret. In particolare sono stati proposti gli algoritmi `R-less-UCB` e `R-ed-UCB`, rispettivamente per lo scenario rising restless e rising rested, fornendo un limite superiore sul regret composto da un termine dipendente dall'istanza del problema e da uno indipendente, che, sotto assunzioni relative al tasso di crescita della funzione di guadagno, è dell'ordine di $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$, dove T è l'orizzonte di apprendimento. Infine, è stata investigata l'efficacia delle soluzioni proposte attraverso un confronto empirico con gli algoritmi dello stato dell'arte per i bandit non-stazionari, facendo uso sia di dati generati in modo sintetico, sia di dati dal mondo reale.

Parole chiave: Multi Armed Bandits, Online Model Selection, Restless Bandit, Rested Bandit, Rising Bandit

Contents

Introduction

Context Real-world problems may require the sequential selection of actions in order to achieve a goal, providing only, at each round, a feedback (aka reward) for the choice. As an example, consider the problem of selecting which dish to order in a restaurant you have never been in. Your first choice is blind, as you do not know how the dish would be prepared there, but if you come back again tomorrow, you will have the knowledge of what you chose the day before and you can decide whether to stick to that option or explore new foods; being very motivated to spot the best available option you come back every evening. Clearly, the process of learning which dish is the best directly impacts your happiness as you have to eat what you ordered, even if it turns out to be really bad. Many other factors may influence the learning, e.g., a cook may be sick and be replaced for some days or a dish may be out of stock. This kind of situation belongs to the so called *online learning* problems. The techniques generally used to cope with such scenarios lie in the field of *stochastic* Multi-Armed-Bandits (MABs), i.e., those methods which are able to learn sequentially (online) exploiting only the feedback observed after an action is performed. A more interesting setting arises when an external factor, such as the time, influences the reward obtained through each action. Consider again the restaurant example, it happens that they have hired an inexperienced young cook. For sure his dishes will improve in quality with days passing. How can we model such setting? Luckily, the classical bandit framework can be extended to deal with *non-stationary* processes, i.e., those problems in which the actions results are influenced by some external factors and suffer a sort of evolution during the agent learning process.

Motivations We want to understand if and how the Multi-Armed-Bandit framework can be applied in the context of *online model selection*, i.e., the task of deciding which algorithm is better at solving a specific task. Consider the case in which there exist many algorithms which can solve a problem, but we do not know which one is better for that specific instance. It is crucial to identify as soon as possible the best solution, not to waste time and computational power feeding bad-performing algorithms. Such a simple scenario hides many pitfalls. For example, consider two binary classification algorithms that are trained with $1,000$ and $2,000$ samples respectively, with the latter achieving a

larger accuracy. The performance of the former is influenced by the less "knowledge" given, as we expect that a larger training set produces a better tuning of the algorithm, improving its performance, but what would it happen if the parts were swapped? Would the former be as good as the latter is now? Would the algorithms reach convergence? When? At which speed? And how can we effectively state which candidate will be better for our problem if we could only assign all our resources to a single one of them? Being motivated by these questions we start our journey.

Goals The goal of this work is to develop and analyze the *rising* Multi-Armed-Bandit setting, a specific non-stationary stochastic bandit framework in which the rewards of each arm are monotonically non-decreasing over time. Specifically, we analyze two different approaches to the reward evolution. On the one hand the **restless** bandit class prescribes a *time*-based evolution, i.e., the reward of an action changes naturally over time. Consider, as an example, the scenario in which some athletes are training for some competitions and a learner has to select which one will partecipate to each event; the performance of the athletes improves naturally over time since they train regularly independently on the fact that the learner decides whether to send them to a race or not. On the other hand the **rested** bandit framework considers a *selection*-based evolution, i.e., the reward of an action changes only when that action is performed. The online model selection example of the two classifiers described earlier clearly lies in the rested bandit class, since the algorithms are trained only when who is in charge of the optimization procedure decides to. We explore the applicability of the rising rested bandits to the online model selection problem. Such an approach is natural since the MAB framework is the common approach for online learning tasks and, generally, we expect learning algorithms to improve their performance the more they are trained, resembling rising evolutions.

Contributions We define the stochastic rising bandit framework, both in the restless and in the rested formulation. We prove that for a rested bandit the assumption of non-decreasing rewards is not enough to ensure the learnability of the problem, hence further assumptions are needed. We propose two algorithms, `R-less-UCB` and `R-ed-UCB`, designed to deal with the restless rising and rested rising settings, respectively. We prove theoretical guarantees on the regret suffered by such algorithm, providing a worst case upper bound on the cumulative expected regret of order $\widetilde{\mathcal{O}}(T^{2/3})$. We provide a wide range of numerical simulations to show the effectiveness of our algorithms w.r.t. state-of-the-art non-stationary bandit algorithms. Finally, we present an online-model-selection experiment on a real-world dataset, showing how `R-ed-UCB` outperforms the benchmarks.

Structure of the Thesis This work is organized as follows. First of all, in Chapter 1 we provide a background for the MAB problem, restating the main concepts of stationary and non-stationary bandits; moving to Chapter 2 we proceed with the description and the formulation of the *rising bandit* problem, a setting in which the non-stationarity of the rewards is assumed to evolve as monotonically non-decreasing functions. In Chapter 3, we focus on the description of the existing techniques which are generally used to solve a MAB problem. Chapter 4 contains the original contribution of this work. We focus on the previously-introduced rising bandit problem, proposing a pair of algorithms, called `R-less-UCB` and `R-ed-UCB`, to operate in the rising restless and rested setting respectively. We provide an in-depth theoretical analysis on both the ideas behind such algorithms and their regret guarantees (Sections 4.2 and 4.1). Finally, Chapter 5 is about the analysis of a set of numerical experiments developed to compare the performances of our solutions with the previously introduced state-of-the-art non-stationary MAB algorithms, using both artificial and real-world data. In Chapter 6 we provide the conclusions.

1 | Background

In this chapter, we first introduce and formalize the classical *stochastic MAB* framework, then we procede analyzing the MAB taxonomy by highlighting the main features of each MAB class. The last part of the chapter is dedicated to the introduction of the main approaches and algorithms generally used to solve a stochastic MAB problem.

1.1. Stochastic MAB

The classical stochastic Multi-Armed-Bandit (MAB) framework [30] models the scenario in which a learner sequentially selects (a.k.a. pulls) some options from a finite set (a.k.a. arms) and receives a feedback (a.k.a. reward) as a consequence of each choice. Consider, as an example, the case in which the owner of a website has to choose which ad to show on the homepage in order to maximize its revenue. This problem can be easily modelled using the MAB framework, where each time a user visits the website the learner chooses an advert from the pool and then observes a positive feedback if the user clicks on the ad or a negative feedback otherwise. Another example of application of the bandit framework is in the so-called dynamic pricing problem, where a company is trying to automatically optimize the price of some product. Customers sequentially arrive and the learner sets the price for each one of them, but the single customer buys the product only if he/she thinks it is convenient w.r.t. his/her valuation. If the product is bought then the learner receives some positive feedback coherent to the price he/she set, or a negative feedback otherwise.

The stochastic MAB framework has been widely applied in a variety of different online-learning applications, such as advertising (Li et al. [33], Nuara et al. [39, 40], Schwartz et al. [50]), routing (Le Ny et al. [31], Parvin and Meybodi [44]), pricing (Sauré and Zeevi [49], Trovò et al. [55]), reccomendation systems (Bresler et al. [13], Li et al. [34]) and medical applications (Aziz et al. [10], Chow and Chang [18], Gittins [24], Thompson [54]).

The earliest reference to the MAB problem dates back to 1933 when Thompson proposed an algorithm which became the basis of many practical approaches today [54]; the bandit

framework however, was formally restated only in 1952 by Robbins, who firstly introduced the notion of regret [47]. We will refer to the Stochastic MAB definition of Lattimore and Szepesvári [30]:

Definition 1.1 (Stochastic MAB [30]). *A K-armed stochastic MAB is a vector of probability distributions $\boldsymbol{\nu} = (\nu_i)_{i\in[K]}$, where $[K] \coloneqq \{1, \dots, K\}$ is the set of available actions.*

The agent and the environment (the bandit) interact sequentially over T rounds (*horizon*), in particular at each round $t \in [T]$:

- the agent selects an arm $I_t \in [K]$ which is fed to the environment;
- the environment samples a reward $R_t \sim \nu_{I_t}$ and reveals it to the learner.

Clearly, a learner would like to select the arm which on average provides the best reward, but in order to spot such an arm the agent may be forced to pull suboptimal arms many times due to the stochasticity of the rewards.

For every arm $i \in [K]$, its *payoff* μ_i is the expectation of the reward from that arm, $\mu_i = \mathbb{E}_{R\sim\nu_i}[R]$; the vector of all payoffs is $\boldsymbol{\mu} = (\mu_i)_{i\in[K]}$ and fully characterizes the bandit. We use the notation i^* to identify the best available arm of the MAB, which is the arm with the largest payoff, $i^* \in \operatorname{argmax}_{i\in[K]} \mu_i$.

When faced with the problem of selecting which arm to pull next, each agent can only exploit the knowledge it has gained while pulling the arms at previous rounds. Such a knowledge is called *history*, identified as the list of the pairs *(arm, reward)* observed up to round t, and is formally defined as $\mathcal{H}_t \coloneqq (I_l, R_l)_{l=1}^{t}$.

Having introduced the concept of history, we can introduce the concept of *policy*:

Definition 1.2 (Policy). *A policy π is a function $\pi : \mathcal{H}_{t-1} \mapsto I_t$, mapping an history to an arm.*

A *policy* completely defines how a learner behaves in every possible situation, hence the term is often used as a synonim of *agent*. To lighten the notation, the abbreviation $\pi(t) \coloneqq \pi(\mathcal{H}_{t-1})$ is widely used, emphasizing even more the response of the learner at each round.

The goal of a learner is to maximize the cumulative profit obtained through its actions up to the horizon T. This quantity is the so-called *cumulative expected reward*:

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

where the expectation is computed over the histories, i.e., the average reward obtained when pulling the arms following the policy π. It is worth noticing that the dependence of $J_{\boldsymbol{\mu}}$ on the actions of the learner does not allow the reduction of a MAB to an optimization problem.

It immediately follows that the *best policy* $\pi^*_{\boldsymbol{\mu},T}$ is the one which achieves the largest cumulative expected reward over the horizon T, or equivalently is the one which always pulls the best available arm i^*:

$$\pi^*_{\boldsymbol{\mu},T} \in \operatorname*{argmax}_{\pi \in [K]^T} J_{\boldsymbol{\mu}}(\pi, T) \implies \pi^*_{\boldsymbol{\mu},T}(t) = i^*.$$

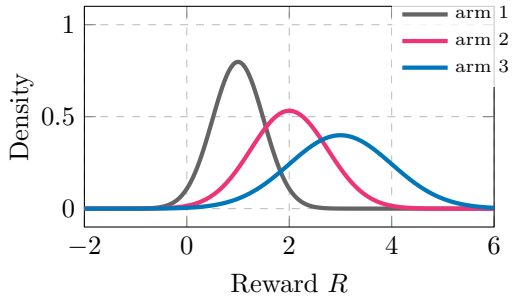
In order to evaluate the performance of a policy π, we can compare it w.r.t. the performance of the optimal policy $\pi^*_{\boldsymbol{\mu},T}$, finding the loss encountered by π as the difference of cumulative expected reward of the two. In the MAB literature such a loss is generally called *regret*.

Definition 1.3 (Cumulative Expected Regret). *The cumulative expected regret of a policy π on a bandit with payoff vector $\boldsymbol{\mu}$ over an horizon T is:*

$$R_{\boldsymbol{\mu}}(\pi, T) \coloneqq J_{\boldsymbol{\mu}}(\pi^*_{\boldsymbol{\mu},T}, T) - J_{\boldsymbol{\mu}}(\pi, T).$$

To practically evaluate the regret of a policy it is useful to rearrange and expand the above definition. Let us consider $N_{i,t} = \sum_{l=1}^{t} \mathbb{1}\{I_l = i\}$, i.e., the number of times arm i has been selected up to round t, it is now possible to decompose the regret over each arm:

$$\begin{aligned}
R_{\boldsymbol{\mu}}(\pi, T) &= J_{\boldsymbol{\mu}}(\pi^*_{\boldsymbol{\mu},T}, T) - J_{\boldsymbol{\mu}}(\pi, T) \\
&= \left[\sum_{t \in [T]} \mu_{i^*}\right] - \mathbb{E}\left[\sum_{t \in [T]} \mu_{I_t}\right] \\
&= [T\mu_{i^*}] - \mathbb{E}\left[\sum_{i \in [K]} N_{i,T}\mu_i\right] \\
&= \mathbb{E}\left[T\mu_{i^*} - \sum_{i \in [K]} N_{i,T}\mu_i\right] \\
&= \mathbb{E}\left[\sum_{i \in [K] \setminus \{i^*\}} N_{i,T}\Delta_i\right] = \sum_{i \in [K] \setminus \{i^*\}} \mathbb{E}\left[N_{i,T}\right]\Delta_i,
\end{aligned}$$



(a) reward distribution of each arm.

$T = 300$	π^*	π_1	π_2
$N_{1,300}$	0	100	75
$N_{2,300}$	0	100	75
$N_{3,300}$	300	100	150
$J_{\boldsymbol{\mu}}(\cdot, 300)$	900	600	675
$R_{\boldsymbol{\mu}}(\cdot, 300)$	0	300	225
$T = 600$	π^*	π_1	π_2
$N_{1,600}$	0	200	150
$N_{2,600}$	0	200	150
$N_{3,600}$	600	200	300
$J_{\boldsymbol{\mu}}(\cdot, 600)$	1800	1200	1350
$R_{\boldsymbol{\mu}}(\cdot, 600)$	0	600	450

(b) performance of policies π_1, π_2.

Figure 1.1: Example of a 3-armed bandit $\boldsymbol{\nu}$.

where $\Delta_i = \mu_{i^*} - \mu_i$ is *the suboptimal gap* for each arm $i \in [K]$, i.e., the difference between the payoff of the best arm and the payoff of arm i.

This formula highlights once again the dependence of the regret on the number of times a suboptimal arm $i \neq i^*$ is pulled, underlining the fact that the more an arm payoff μ_i is lower than the best arm payoff μ_{i^*}, the less that arm should be pulled.

Example 1.1. *Let us consider a* 3*-armed bandit* $\boldsymbol{\nu}$ *where each arm reward is sampled from a Gaussian distribution:* $\nu_1 \sim \mathcal{N}(1, 0.5)$, $\nu_2 \sim \mathcal{N}(2, 0.75)$, $\nu_3 \sim \mathcal{N}(3, 1)$ *(Figure 1.1a). Being the payoffs* $\mu_3 > \mu_2 > \mu_1$, *arm* 3 *is clearly the best available option for a learner. However the stochasticity (noise) of the process does not allow the learner to observe each payoff* μ_i *directly. Consider two different strategies* π_1 *and* π_2, *the former selects each arm an equivalent number of times, while the latter pulls the best arm* 50% *of the times and shares the remaining pulls equally between the other arms. We evaluated the performance of each of the two policies over two different time horizons* $T_1 = 300$ *and* $T_2 = 600$ *reporting the results in Table 1.1b. Considering the total number of pulls of each arm* $N_{i,T}$ *over the horizon, the cumulative expected reward is* $J_{\boldsymbol{\mu}}(\pi, T) = \sum_{i \in \{1,2,3\}} N_{i,T}\mu_i$; *recalling that the optimal policy* π^* *is the one which always pulls arm* 3, *we can compare the cumulative expected reward of our two policies with the one of the best policy, in order to find out their cumulative expected regret (Definition 1.3). In the following, we will provide the tools to state whether the performance of a policy is good or bad, for now it is sufficient to notice that for both the policies the regret increases linearly with the time horizon* T.

1.2. MAB taxonomy

The stochastic MAB framework is very general and its application is possible in many different problems. For example, in the advertisment problem presented before, the owner of the website may notice that a returning user is less likely to click again on the ad he was shown the previous time. Once again the MAB framework can model this situation since it imposes no constraints on the reward distributions shape or stationarity. It is clear that an algorithm which is tuned to solve a problem in which the reward distributions of each arm are stationary in time will behave differently if a sort of evolution of the arms rewards is allowed. Different situations and assumptions also lead to different theoretical guarantees on the regret, hence the need of a subclassification of the general MAB framework.

Stationary and Non-Stationary Bandits Among the family of stochastic MABs it is possible to distinguish between two different categories: the *stationary stochastic* bandits and the *non-stationary stochastic* ones. The former includes all the bandit problems in which the reward distributions are fixed and do not evolve during the learning process, i.e., for each $i \in [K], t \in [T]$, each distribution ν_i is equal to a fixed distribution c_i and clearly its payoff $\mu_i = \mathbb{E}[c_i]$ is constant; this is the case of the *classical* stochastic MAB setting and of Example 1.1 presented earlier. In literature, addressing a problem simply as "stochastic MAB" often understate the stationarity assumption. On the other hand the non-stationary family includes all the MAB problems in which the reward distributions associated to each arm evolve during the learning process. Among this class it is possible to distinguish between *restless* and *rested* MABs.

Restless and Rested Bandits The *restless* and *rested* bandit settings have been introduced in the context of non-stationary bandits by Tekin and Liu [53] and further developed by Ortner et al. [41] and Russac et al. [48] in the restless version, and by Mintz et al. [38] and Pike-Burke and Grunewalder [45] in the rested one. The evolution of the payoff was originally modeled via a suitable process, e.g., a Markov chain with finite state space, while recently, the terms restless and rested have been employed to denote arms whose payoff changes as time passing, or whenever being pulled, respectively (Seznec et al. [51, 52]). Restless bandits model many practical situations in which the arms' payoff may change over time due to intrinsic modifications of the arms or of the environment. In this setting the reward evolution can be described by a function of the round t, i.e., $\nu_i(t)$. Through this dependency it is possible to model environments in which the payoffs evolve because of some external factors that the learner cannot influence. An example of this scenario can be found in the stock selection problem where an agent would like to create

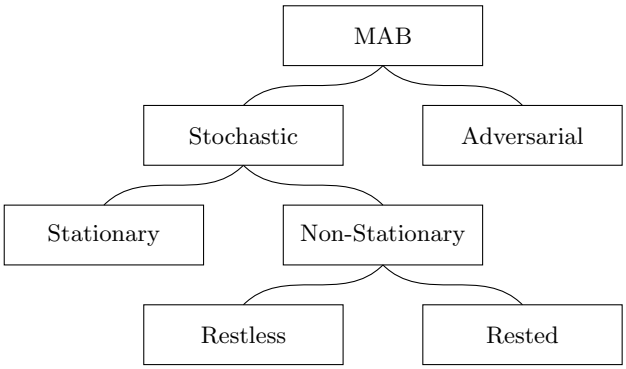


Figure 1.2: Multi-Armed-Bandit taxonomy.

his portfolio selecting some stocks. Clearly the stock market is not stationary; moreover each stock price changes as time passes and the learner cannot control it in any way. On the other hand, while in restless MABs the arms' payoff changes *naturally* over time, a different setting arises when the payoff changes as an effect of *pulling* the arm. This situation is described in the rested bandits framework, in which the actions of the agent directly trigger the evolution of the arms' reward distributions. In such a scenario, the reward evolution is described by a function of the number of times an arm has been pulled, i.e., $\nu_i(N_{i,t})$. An example of this setting is the task of allocating computational resources for online model selection, we would like to know which algorithm is more promising than the others in order to give it more computational power, improving the performance of that model but at the same time slowing the training of our other candidates.

Adversarial MAB A dual approach to the stochastic MAB framework is the *adversarial* MAB, which was firstly introduced by Auer et al. [7]. Such a model is used in situations in which the rewards cannot be represented through stochastic distributions but are considered to be the result of an interaction with an external adversary. Hence no assumption on the nature of the reward is required. In principle, it is possible to use the adversarial approach to take into account the non-stationarity of some situations, however, in practice, the performance obtained is unsatisfactory because the non-stationarity of real-world cases is far from being adversarial.

Figure 1.2 summarizes the taxonomy of MAB models presented above, helping claryfing the distinctions and subclassifications.

1.3. Algorithms for Stationary MABs

In this Section, we introduce three of the mostly used stationary bandit algorithms, with a particular focus on the main ideas over which each algorithm is built. As we shall see in Chapter 3, the knowledge of stationary bandit algorithms is crucial to successfully develop proper solutions for the non-stationary MAB setting.

1.3.1. Regret Bounds

Lai and Robbins [29] provided a *lower bound* analysis on the regret obtained by a generic learning agent, showing that no policy can achieve a cumulative expected regret lower than a minimum quantity which depends on the arms payoff values:

Theorem 1.1 (Regret Lower Bound [29]). *In a stochastic MAB $\boldsymbol{\nu}$ with Bernoulli distributed rewards, for every policy π that satisfies $\lim_{T\to+\infty} T^{-1} J_{\boldsymbol{\mu}}(\pi, T) = \mu_{i^*}$, then:*

$$\lim_{T\to+\infty} R_{\boldsymbol{\mu}}(\pi, T) \geq \log T \sum_{i\in[K]\setminus\{i^*\}} \frac{\Delta_i}{\mathcal{KL}(\mu_i, \mu_{i^*})},$$

where $\mathcal{KL}(x, y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$ is the Kullback-Leibler divergence (a.k.a. relative entropy).

This result is important since it offers a benchmark to the performance of any MAB algorithm. The closer the obtained regret to this bound, the better the algorithm. However, in practice, the main focus of any bandit study is to guarantee an *upper bound* on the worst case regret, in order to achieve a performance which satisfy the *no-regret* property:

Definition 1.4 (No-Regret Property). *If for any bandit $\boldsymbol{\nu}$, the regret obtained by the policy π satisfies:*

$$R_{\boldsymbol{\mu}}(\pi, T) \leq \widetilde{\mathcal{O}}\left(T^{\alpha}\right), \quad \text{with } \alpha < 1,$$

where $\widetilde{\mathcal{O}}$ disregards logarithmic terms, then the policy π is no-regret.

A learning agent which does not satisfy Definition 1.4 may achieve a regret which grows linearly with the horizon of the problem, meaning that he is not learning effectively. It is worth noticing that a policy π may satisfy Definition 1.4 for a specific class of bandits only (e.g., stationary bandits), which is what often happens in practice.

1.3.2. Optimistic Exploration for Stochastic MABs

The key concept behind the firsts (stationary) stochastic MAB algorithms is the so called *optimism in face of uncertainty* approach, firstly introduced by Lai and Robbins [29]. Imagine visiting a city in a foreign country and having to choose where to eat, we can stick to some well known fast-food chain which for sure will grant us a fair lunch, or explore an unknown local restaurant which may turn out to be the best one we have ever been in. Giving a chance to the local restaurant is an optimistic approach to the problem. Once we have tried the new restaurant a few times, we can update our ranking and make a final decision. This optimistic approach encourages the exploration of unknown options. Consider instead a pessimistic approach: we would never try the local restaurant and possibly never find out that is better than the fast-food we are going to, leading to a regret linearly increasing.

The application of the optimistic principle to a bandit is generally realized assigning to each arm a quantity, called *Upper Confidence Bound (UCB)* that is, with high probability, an overestimate of the unknown payoff of the arm μ_i. This quantity depends both on the average reward obtained by pulling the arm and the number of times the arm has been pulled. Intuitively, the more the arm is pulled, the more the UCB will be a strict overestimation of the average reward of the arm, while on the other hand, the less the arm is pulled, the looser the bound.

Intuitevly, we can understand how the regret sublinearity is achieved: at each round t we selct the arm I_t which has the highest UCB and consequently update its bound using the newly observed data, thightening it. If we keep pulling the same arm, at some point the confidence bound will be roughly equivalent to the real payoff; but this can happen only if the arm we chose is the best one since otherwise at some point another arm's UCB would be higher than our arm's payoff. On the other hand the exploration is encouraged by the optimistic approach, since sooner or later an arm will be explored because it has a UCB higher than the others'.

UCB

The first application of this concept to an algorithm for stochastic MABs was by Lai [28] who proposed in 1987 the first version of a UCB algorithm; later on Katehakis and Robbins [26] studied the application of an upper confidence bound to Gaussian distributed rewards, while Agrawal [3] proposed the usage of the samples' average as the core of the upper bound. We will refer to the algorithm called `UCB1` proposed by Auer et al. [8] since it comes with theoretical guarantees on the regret when the payoffs are bounded in $[0, 1]$

and the rewards are 1/2-subgaussians.

Definition 1.5 (Subgaussian Distribution). *A random variable $R \sim \nu$ is σ^2-subgaussian if for every $\lambda \in \mathbb{R}$:*

$$\mathbb{E}_{R\sim\nu}[e^{\lambda(R-\mu)}] \leq e^{\frac{\sigma\lambda^2}{2}}, \quad \text{where } \mu = \mathbb{E}_{R\sim\nu}[R].$$

Intuitevly the tails of the distributions are dominated by the tails of a Gaussian with σ^2 as variance.

The upper confidence bound used in `UCB1` (Algorithm 1.1) relies on the average of the samples obtained so far from each arm $i \in [K]$:

$$\widehat{\mu}_i(t) = \frac{1}{N_{i,t}} \sum_{l=1}^{t} R_l \mathbb{1}\{I_l = i\} \tag{1.1}$$

In addition to the samples average $\widehat{\mu}_i(t)$, the bound is obtained considering an extra exploration bonus $\sqrt{\frac{2\log t}{N_{i,t}}}$ which grows with time. This is particularly important in order to obtain an optimistic estimate of the next reward from arm i. The direct dependence on t makes the exploration factor increase at each round t regardless arm i is pulled, encouraging exploration; on the other hand the more arm i is pulled, i.e., the greater $N_{i,t}$, the less the bonus factor and the accurate the samples' average $\widehat{\mu}_i$.

Algorithm 1.1 `UCB1`

Input: T horizon, K arms.

1: Pull each arm once.
2: **for** $t = K+1, \ldots, T$ **do**
3: Pull arm $I_t \in \text{argmax}_{i\in[K]} \left\{\widehat{\mu}_i(t) + \sqrt{\frac{2\log t}{N_{i,t}}}\right\}$, where $\widehat{\mu}_i$ is defined in Equation (1.1)
4: **end for**

Theorem 1.2 (`UCB1` Regret [8]). *In a K-armed bandit with reward distributions bounded in $[0,1]$, the regret suffered by `UCB1` is:*

$$R_{\boldsymbol{\mu}}(\texttt{UCB1}, T) \leq 8 \sum_{i\in[K]\setminus\{i^*\}} \left(\frac{\log(T)}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{j\in[K]} \Delta_i\right).$$

`UCB1` not only is an effective strategy to solve a stationary stochastic MAB problem, but it also describes a more general approach which can be extended and used in many others

MAB contexts.

KL-UCB

A straightforward evolution of `UCB1`, and one of the best results in terms of theoretical guarantees on the regret achieved in the context of stationary MABs, is represented by the `KL-UCB` algorithm [21]. Garivier and Cappé showed that in the particular case of bandits with *Bernoulli* distributed rewards, i.e., $\nu_i \sim \mathcal{B}(\mu_i)$ for each arm $i \in [K]$, `KL-UCB` reaches the regret lower bound of Theorem 1.1. `KL-UCB` can also be used in the general case of continuously distributed rewards, however it is a necessary assumption that every reward R_t is bounded in $[0, 1]$. For simplicity the algorithm is hereby presented considering Bernoulli distributed rewards.

Algorithm 1.2 `KL-UCB`

Input: T horizon, K arms, $c \in \mathbb{R}_{\geq 0}$.

1: **for** $t = 1, \dots, K$ **do**
2: Pull arm $I_t = t$
3: $S_t = R_t$
4: **end for**
5: **for** $t = K + 1, \dots, T$ **do**
6: Pull arm $I_t \in \operatorname{argmax}_{i \in [K]} \left\{ \max_{q \in [0,1]} \left\{ N_{i,t} \cdot \mathcal{KL}\left(\frac{S_i}{N_{i,t}}, q\right) \leq \log t + c \log\log t \right\} \right\}$
where $\mathcal{KL}(x, y) = x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$
7: $S_{I_t} = S_{I_t} + R_t$
8: **end for**

Notice the similarities between `KL-UCB` and `UCB1`. After the initialization step in which all the arms are pulled once, the only difference is the structure of the upper confidence bound through which the arm to pull is selected.

For completeness, we report here the regret guarantees of `KL-UCB`, which, as previously mentioned, matches the lower bound performance of Lai and Robbins (Theorem 1.1).

Theorem 1.3 (`KL-UCB` Regret [21]). *In a K-armed bandit with independent rewards bounded in $[0, 1]$, the cumulative expected regret of Algorithm 1.2 with $c = 3$ satisfies:*

$$\limsup_{T \to +\infty} R_{\boldsymbol{\mu}}(\textit{KL-UCB}, T) \leq \log T \sum_{i \in [K], i \neq i^*} \frac{\Delta_i}{\mathcal{KL}(\mu_i, \mu_{i^*})}.$$

1.3.3. Bayesian Approaches for Stochastic MABs

The dual counterpart of the optimistic principle is represented by the Bayesian approach: a set of probability distributions is used to choose which arm to pull based on its "probability" of being the best arm; at each round the newly collected sample is then used to update such distributions.

Thompson Sampling

The `TS` algorithm [54], represents the most famous Bayesian solution for the stochastic MAB problem. The algorithm initially exploits the *prior* knowledge on the reward distributions in order to choose the arm to pull, then it procedes updating and refining such knowledge based on the outcome of the pull. What is generally used is a pair of *conjugate priors*, i.e., a pair of probability distributions which can be easily combined, in order to perform the update step. `TS` relies on the assumption of Bernoulli distributed rewards and it uses *Beta* distributions to represent the knowledge of the arms, since they can be easily updated when the outcome of the pull is in $\{0, 1\}$. It is possible to extend the algorithm in order to deal with bandits with rewards bounded in the interval $[0, 1]$, operating a reduction to Bernoulli bandits by sampling at each round from the observation R_t a bernoullian value $X_t \sim \mathcal{B}(R_t)$ and using the outcome X_t to update the estimators. However, we present here the algorithm originally designed for Bernoulli rewards. Such algorithm considers no prior knowledge on the distributions, i.e., uses uniform distributions as priors (remind that $\mathcal{U}(0, 1) \equiv \mathcal{B}eta(1, 1)$).

Algorithm 1.3 `Thompson Sampling (TS)`

Input: T horizon, K arms.

For each arm $i \in [K]$ set $S_i = 0$, $F_i = 0$. // Successes(1) and Failures(0) from i
for $t = 1, \dots, T$ **do**
  For each arm $i \in [K]$, sample $\theta_i \sim \mathcal{B}eta(S_i + 1, F_i + 1)$.
  Pull arm $I_t \in \text{argmax}_i \theta_i$ and observe reward R_t.
  if $R_t = 1$ **then**
    $S_{I_t} = S_{I_t} + 1$
  else
    $F_{I_t} = F_{I_t} + 1$.
  end if
end for

The original `TS` algorithm was introduced by Thompson in 1933, however its regret analy-

sis was not provided in the original work; we report here the results obtained by Agrawal and Goyal [4]:

Theorem 1.4 (TS Regret [4]). *For the K-armed stochastic bandit problem, Algorithm 1.3 has cumulative expected regret:*

$$R_{\boldsymbol{\mu}}(\mathtt{TS}, T) \leq \mathcal{O}\left(\left(\sum_{i \in [K] \setminus \{i^*\}} \frac{1}{\Delta_i^2}\right) \log T\right).$$

It is worth noticing that in many practical cases TS often achieves better performances w.r.t. UCB algorithms.

2 | Rising Bandits

In this chapter, we provide a formal definition of the non-stationary multi-armed-bandit, with a particular focus on the restless and rested MAB setting, recalling the definitions provided by Seznec et al. [52]. We then proceed formalizing the main aspects of the *rising bandit* problem, discussing the main assumptions under which learning is possible in this setting.

2.1. Problem Formulation

Recall from Chapter 1 the definition of a multi-armed-bandit with K arms:

Definition 1.1 (Stochastic MAB [30]). *A K-armed stochastic MAB is a vector of probability distributions $\boldsymbol{\nu} = (\nu_i)_{i \in [K]}$, where $[K] \coloneqq \{1, \dots, K\}$ is the set of available actions.*

While in a stationary bandit the probability distribution ν_i associated to each arm is fixed, in a *non-stationary* context we can consider each one as a function $\nu_i : \mathbb{N}^2 \to \Delta(\mathbb{R})$ which depends on a pair of parameters $(t, n) \in \mathbb{N}^2$, where $\Delta(\mathbb{R})$ denotes the set of probability distributions over $\mathbb{R}$. At each round $t \in [T]$, the reward is sampled from $\nu_{I_t}(t, N_{I_t,t-1})$, thus, the observed reward depends, in general, on the current round t and on the number of pulls $N_{I_t,t-1}$ of arm I_t performed so far. Being the payoffs the expected value of the distributions, they are functions $\mu_i : \mathbb{N}^2 \to \mathbb{R}$ themselves, defined as $\mu_i(t, n) = \mathbb{E}[\nu_i(t, n)]$.

Restless and Rested arms Restless and Rested MAB were firstly defined by Tekin and Liu [53], however the presence of an underline Markov Chain is not suitable for our work since in a non-decreasing reward setting every state of the chain can only be visited once. We will refer to the definition of Seznec et al. [52]:

Definition 2.1 (Restless and Rested Arms). *Let $\boldsymbol{\nu}$ be a MAB and let $i \in [K]$ be an arm, we say that:*

- *i is a* restless *arm if, for every round $t \in [T]$ and number of pulls $n \in \mathbb{N}$, we have $\mu_i(t, n) = \mu_i(t)$;*
- *i is a* rested *arm if, for every round $t \in [T]$ and number of pulls $n \in \mathbb{N}$, we have $\mu_i(t, n) = \mu_i(n)$.*

A restless arm is in all regards a non-stationary arm [11], and it is suitable for modeling a natural phenomenon that evolves as time passes, independently of the agent intervention. On the other hand, the payoff of a rested arm changes when being pulled and, therefore, it models phenomena that evolve as a consequence of the agent intervention.

Definition 2.2 (Restless and Rested MAB). *A K-armed bandit is* restless *(resp.* rested*) if all of its arms are restless (resp. rested).*

Rising Bandits The goal of the rising bandit framework is to represent a scenario in which the payoffs are monotonically non-decreasing over time (or pulls, depending on the bandit being restless or rested); such model was firstly introduced by Heidari et al. [25] in its deterministic version as a dual problem to the rotting bandits, in which the payoffs are assumed to be monotonically non-increasing. However as Heidari et al. [25] showed, it represents a significantly more complex problem, even for deterministic arms, and cannot be addressed with the same approaches. Indeed, in the non-decreasing payoff setting, an assumption commonly employed is the *concavity* of the payoff functions (Li et al. [35]). We revise the *improving* bandits definition introduced by Heidari et al. [25] to identify such bandits with non-decreasing concave payoffs, later denoted as *rising* bandits in Li et al. [35]:

Assumption 2.1 (Non-Decreasing Payoffs). *Let $\boldsymbol{\nu}$ be a MAB, for every arm $i \in [K]$, number of pulls $n \in \mathbb{N}$, and round $t \in [T]$, functions $\mu_i(\cdot, n)$ and $\mu_i(t, \cdot)$ are non-decreasing. In particular, we define the* increments*:*

$$\begin{aligned} &\textit{Restless arm:} && \gamma_i(t) \coloneqq \mu_i(t+1) - \mu_i(t) \geq 0, \\ &\textit{Rested arm:} && \gamma_i(n) \coloneqq \mu_i(n+1) - \mu_i(n) \geq 0. \end{aligned}$$

From an economic perspective, $\gamma_i(\cdot)$ represents the *increase of total return* (payoff) we experience by adding a factor of production, i.e., letting time evolve for one unit (restless) or pulling the arm (rested).

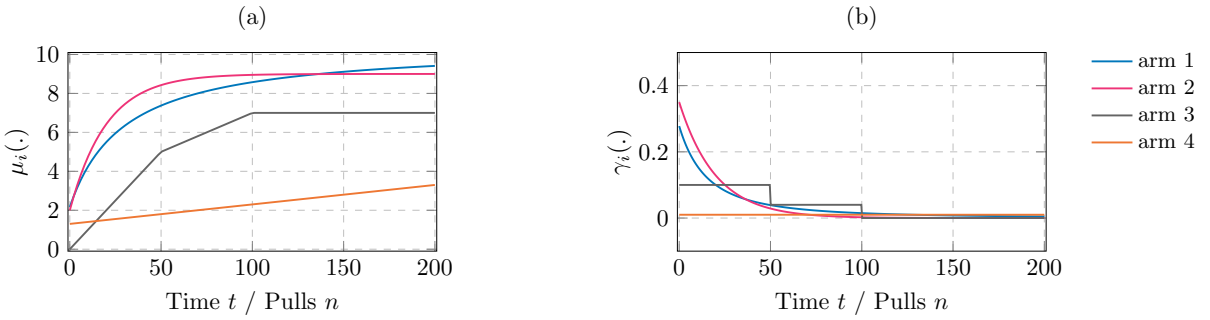


Figure 2.1: Example of a rising bandit: payoff evolution (a), growth evolution (b)

Assumption 2.2 (Concave Payoffs). *Let $\boldsymbol{\nu}$ be a MAB, for every arm $i \in [K]$, number of pulls $n \in \mathbb{N}$, and round $t \in [T]$, functions $\mu_i(\cdot, n)$ and $\mu_i(t, \cdot)$ are concave, i.e.:*

$$\begin{aligned} &\textit{Restless arm:} \quad \gamma_i(t+1) - \gamma_i(t) \leq 0, \\ &\textit{Rested arm:} \quad \gamma_i(n+1) - \gamma_i(n) \leq 0. \end{aligned}$$

As noticed by Heidari et al. [25], the concavity assumption corresponds, in economics, to the *decrease of marginal returns* that emerges when adding a factor of production, i.e., letting time evolve for one unit (restless) or pulling the arm (rested).

Definition 2.3 (Rising MAB). *We define* rising *a stochastic MAB in which both Assumption 2.1 and Assumption 2.2 hold.*

For the seek of simplicity and without loss of generality, we assume that the payoffs are bounded in $[0, 1]$, and that the rewards are σ^2-subgaussian random variables (Definition 1.5)

Example 2.1. *Consider a bandit with $K = 4$ arms, a problem horizon $T = 200$, and assume that the rewards are deterministic realizations, i.e., no stochasticity is considered; this bandit "shape" is defined in Figure 2.1a and in Figure 2.1b, which shows the evolution of the arms payoffs μ_i over time (or pulls) and the evolution of the growths γ_i, i.e., the relative step-by-step increase of the payoff, respectively. First of all from Figure 2.1a, it is possible to notice how all the 4 functions trivially satisfy Assumption 2.1, being the payoffs non-decreasing functions. Moreover, from Figure 2.1b, which shows how the derivative of the payoff functions are decreasing and progressively approaching 0, we can state that the concavity assumption (Assumption 2.2) is satisfied. Hence the one presented in Figure 2.1 is a rising bandit. This example shows how different can the reward evolution be, and how the growths directly impact a learning process.*

In Chapter 3, we will present the main works related to the rising bandit problem and

the main techniques used in the field of non-stationarity bandits which can be applied to the rising bandit setting. In Chapter 4, we will introduce and analyze learning algorithms for both restless (Section 4.2) and rested (Section 4.1) rising bandits. We will present *optimistic* algorithms to cover both cases, analyzing how to create an upper bound in the different settings, and provide theoretical guarantees on the cumulative expected regret.

3 | Literature Analysis

The first part of the chapter is about the main works related to the rising bandit framework, and particularly focus on the model selection problem. In the second part of the chapter, we provide the description of the main techniques used in the field of non-stationary stochastic MAB problems, we present the main state-of-the-art algorithms for non-stationary bandits underlining their assumptions and regret guarantees. Those algorithms will be used as benchmark for our numerical simulations of Chapter 5.

3.1. Related Bandit Frameworks

As previously introduced in Chapter 2 the distinction between *restless* and *rested* bandit problems is a key part of our work, hence the need to study the existing literature on non-stationary bandits taking into account both approaches. Indeed, it is clear that algorithms which address the former setting can be used in the latter only under certain circumstances or when tuned accordingly, and viceversa. The majority of the works discuss algorithms which are designed to deal with a restless environment, often generically addressed as non-stationary bandits. Compared to the rested setting, there exist algorithms designed for specific situations, i.e., in known reward shapes or noiseless settings.

In this section, we highlight the bandit works which are mostly related to the rising bandit scenario and to the online model selection problem.

Improving Bandits As previously mentioned in Chapter 2, rising bandits applied to the online model selection problem, have been already tackled in a deterministic version by Heidari et al. [25] and Li et al. [35], under the name of *improving bandits*. The former work discusses an online algorithm to minimize the regret when considering the task of selecting an increasing concave function among a finite set, under the assumption that the learner receives a feedback about the true value of the reward function, i.e., no stochasticity is considered, while the latter discusses the problem of parameter optimization for machine learning models proposing once again an efficient algorithm for deterministic rising MABs. Stochasticity for rising bandits was firstly introduced by Cella et al. [16] under

the assumption that the payoffs evolve according to a specific family of functions known a priori to the learner, whose task become the correct estimation of the parameters of such function according to the observations. The need for knowing the parametric form of the payoff makes these approaches hardly applicable for arbitrary increasing functions.

Corralling Bandits An approach to the online model selection problem is respresented by *corralling bandits* (Abbasi-Yadkori et al. [1], Agarwal et al. [2], Arora et al. [6], Pacchiano et al. [42, 43]), whose goal is to try to minimize the regret of a process choosing among a finite set of bandit algorithms. This setting figures a meta-bandit operating on the so-called base algorithms, and is characterized by the following assumptions. First of all, each arm must correspond to a learning algorithm, the base algorithm, which itself operates on a bandit; moreover, each base algorithm must be endowed with a (possibly known) regret bound; sometimes additional conditions (such as stability) are required.

Recharging Bandits The *recharging bandits* designed by Kleinberg and Immorlica [27] represent a framework which assumes that the reward of an arm decreases when it is pulled but increases when it is not pulled, modeling phenomena in which continuously performing the same choice is counterproductive. Differently from rising bandits, the recharging bandits model only assumes local monotonicity occurring between two consecutive pulls, indeed the expected reward is neither globally rising nor rotting and it generally depends on the amount of time passed from the last time we selected an arm.

Rotting Bandits The *rotting bandits* setting was introduced by Levine et al. [32]. This framework models the case in which the performance of the arms degradates the more they are pulled; such a setting may arise in many real-world problems, e.g., advertisment, reccomendation and networking. Knowing the monotonicity property of the rewards allows deriving more specialized algorithms, exploiting the characteristics of the underlying process and decreasing the regret w.r.t. unrestricted cases. Due to the duality w.r.t. the rising bandit problem, and being one of the few example of a rested MAB, we report here the main assumption behind the rotting bandits:

Assumption 3.1 (Non-Increasing Payoffs)**.** *In a rotting MAB the arms' payoffs satisfy:*

$$\mu_i(N_{i,t}) \geq \mu_i(N_{i,t}+1) \quad \forall i \in [K], \forall t \in [T].$$

Abruptly Changing Bandits This scenario refers to the class of non-stationary bandits for which the reward evolution is modeled as piece-wise stationary functions, i.e., the

reward values are stationary except for a finite number of rounds (called *breakpoints*) in which they assume a different, stationary, value. The approaches to this setting can be diveded in two families. Passive methods (Garivier and Moulines [22], Trovò et al. [56]) exploit only the most recent observations to predict the future outcome. Active methods (Besson et al. [12], Cao et al. [15], Liu et al. [36]) try to identify the change and adapt the decision policy accordingly. Those are the so called *change detection* (CD) methods. Advanced CD-MAB algorithms have been developed as well, to take into account the possible regularities in the non stationary process (Re et al. [46]). Unfortunately these approaches cannot be used for a rising bandit problem in which, theoretically, the number of breakpoints is equal to the horizon T, loosening the regret guarantees of such algorithms.

3.2. Related Bandit Algorithms

Due to the already discussed unbalance favouring the more common restless scenario, in this section we report five algorithms used in the context of restless MABs. The first three, `SW-UCB` [23], `SW-KL-UCB` [19] and `SW-TS` [56], are the non-stationary counterpart of the already discussed stationary bandit algorithms, while the last two, `Rexp3` [11] and `Ser4` [5], represent a different approach to the problem. Compared to the rested MABs we present the case of rotting bandits, discussing the `RAW-UCB` [52] algorithm.

3.2.1. Algorithms for Restless MABs

The *restless MAB* problem is generally addressed both using passive methods (e.g., Garivier and Moulines [23], Auer et al. [9], Trovò et al. [56], Besbes et al. [11]) and active ones (e.g., Liu et al. [36], Besson et al. [12], Cao et al. [15]). The former algorithms base their selection criterion on the most recent feedbacks, while the latter actively try to detect if a change in the arms' rewards occurred in order to use only data gathered after the last change. Garivier and Moulines [22, 23] discussed a sliding-window approach to solve the problem assuming abruptly-changing payoffs, while Trovò et al. [56] analyzed in their work smoothly-changing payoffs too, proposing a sliding-window Thompson Sampling approach. Besbes et al. [11] studied the problem considering bounded reward variations, i.e., a maximum variation budget allowed to the arms' payoffs, proposing an algorithm based on the Adversarial MAB framework; many other works assumed the knowledge of the reward evolution function shape, such as Cella et al. [16], or lay on some monotonicity assumptions (i.e., Kleinberg and Immorlica [27], Pike-Burke and Grunewalder [45]).

SW-UCB

The `Sliding-Window-UCB` (`SW-UCB`) algorithm [23] exports the Upper Confidence Bound principles (recall Section 1.3.2) to a non-stationary context. In order to achieve such a result the sample means of the rewards used in `UCB1` to bound future rewards, are now modified in order not to consider all the observations gained from the start of the learning process but only the last τ ones, where τ is a *sliding window* whose size is fixed a priori.

The algorithm exploits the following estimators and parameters:

$$N_t(\tau, i) = \sum_{l=t-\tau+1}^{t} \mathbb{1}\{I_l = i\}, \tag{3.1}$$

$$\widetilde{X}_t(\tau, i) = \frac{1}{N_t(\tau, i)} \sum_{l=t-\tau+1}^{t} R_l \mathbb{1}\{I_l = i\}, \tag{3.2}$$

$$c_t(\tau, i) = \sqrt{\frac{\xi \log(\min(t, \tau))}{N_t(\tau, i)}}, \tag{3.3}$$

where Equation (3.1) represents the weighted number of times arm i has been pulled, considering recent pulls more important, Equation (3.2) defines the weighted average of the observation from arm i, considering only the one that are more recent and fits in the sliding window; finally Equation (3.3) defines a constant factor c to be used as an exploration bonus.

Algorithm 3.1 `SW-UCB`

Input: T horizon, K arms, τ window-size, $\xi \in [0, 1]$.

1: **for** $t = 1, \ldots, K$ **do**
2: Pull arm $I_t = t$
3: **end for**
4: **for** $t = K + 1, \ldots, T$ **do**
5: Pull arm $I_t \in \text{argmax}_{i \in [K]} \left\{ \widetilde{X}_t(\tau, i) + c_t(\tau, i) \right\}$,
where $\widetilde{X}_t$, c_t are defined in Equations 3.2, 3.3.
6: **end for**

The main assumption under which the theoretical results are achieved is the abruptly changing environment, meaning that the payoffs are stationary over time with the exception of some rounds (a.k.a. breakpoints) in which they drastically change and then become stationary again with those new values, representing full-fledged piece-wise sta-

tionary functions.

Theorem 3.1 (SW-UCB Regret [23]). *Let Υ_T be the number of breakpoints, then for $\xi > 0.5$ and $h = 2\sqrt{T \log T / \Upsilon_T}$:*

$$\mathbb{E}\left[\widetilde{N}_T(i)\right] \leq \mathcal{O}\left(\sqrt{\Upsilon_T T \log T}\right),$$

where $\widetilde{N}_T(i)$ represents the number of times Algorithm 3.1 pulls arm i when it is not optimal.

Theorem 3.1 requires the prior knowledge of the number of abrupt changes, however similar results have been obtained by Auer et al. [9] without requiring the knowledge of the number of breakpoints.

SW-KL-UCB

Combes and Proutiere [19] proposed the application of a sliding-window of size h to some stationary MAB algorithms in order to use those solutions in a non-stationary context. We will consider here the sliding-window version of Algorithm 1.2, called `SW-KL-UCB`.

Let's introduce the following quantities:

$$\widetilde{N}_i(t) = \sum_{l=t-h}^{t} \mathbb{1}\{I_l = i\}, \tag{3.4}$$

$$\widetilde{S}_i(t) = \sum_{l=t-h}^{t} R_l \mathbb{1}\{I_l = i\}, \tag{3.5}$$

where Equations (3.4) and (3.5) represent respectively the number of times arm i has been pulled and the sum of observations from arm i in the last h rounds.

Algorithm 3.2 `SW-KL-UCB`

Input: T horizon, K arms, h window-size, $c \in \mathbb{R}_{\geq 0}$.

1: **for** $t = 1, \ldots, K$ **do**
2: Pull arm $I_t = t$ and set $\tilde{N}_t = 1$, $\tilde{S}_t = R_t$
3: **end for**
4: **for** $t = K+1, \ldots, T$ **do**
5: Pull arm $I_t \in \operatorname{argmax}_{i \in [K]} \left\{ \max_{q \in [0,1]} \left\{ \tilde{N}_i \cdot \mathcal{KL}\left(\frac{\tilde{S}_i}{\tilde{N}_i}, q\right) \leq \log t + c \log\log t \right\} \right\}$
where $\mathcal{KL}(x, y) = x \log(x/y) + (1-x)\log((1-x)/(1-y))$
6: Update $\tilde{N}_i$, $\tilde{S}_i$ for each $i \in [K]$
7: **end for**

A fundamental assumption which must be taken into account when using the SW-KL-UCB algorithm is that the arms' rewards must be Bernoulli distributed. It is however possible to consider bandits with rewards bounded in $[0,1]$ by performing a sampling step as already discussed in Section 1.3.3.

Theorem 3.2 (SW-KL-UCB Regret [19]). *In a K-armed bandit with independent rewards bounded in $[0,1]$, the cumulative expected regret of Algorithm 3.2 with $c = 3$ satisfies:*

$$\limsup_{T\to+\infty} R_{\boldsymbol{\mu}}(\textit{SW-KL-UCB}, T) \leq \mathcal{O}\left(\sqrt{T}\log T\right)$$

SW-TS

Trovò et al. [56] proposed a sliding-window approach to the TS algorithm. As already described in Section 1.3.3 the assumption of Bernoulli distributed rewards can be dropped introducing an extra sampling step, however for coherence we will present the algorithm considering Bernoulli distributed rewards. Let us introduce the following quantities:

$$\widetilde{N}_{i,t} = \sum_{l=\max\{t-h+1,1\}}^{t} \mathbb{1}\{I_t = i\}, \tag{3.6}$$

$$\widetilde{S}_{i,t} = \sum_{l=\max\{t-h+1,1\}}^{t} R_l \mathbb{1}\{I_t = i\}, \tag{3.7}$$

where Equations (3.6) and (3.7) represent respectively the number of times arm i has been pulled and the cumulative reward collected from arm i in the last $\min(h,t)$ rounds.

Algorithm 3.3 SW-TS

Input: T horizon, K arms, h window-size, $\{r_{i,0}\}_{i\in[K]}$ prior distributions.

1: **for** $t = 1,\dots,T$ **do**
2: **for all** $i \in [K]$ **do**
3: Compute $r_{i,t} = \mathcal{B}eta(\tilde{S}_{i,t} + 1, \tilde{N}_{i,t} - \tilde{S}_{i,t} + 1)$
4: Sample $\theta_{i,t} \sim r_{i,t}$
5: **end for**
6: Pull arm $I_t \in \operatorname{argmax}_{i\in[K]} \theta_{i,t}$, observe R_t and update $\tilde{N}_{i,t}$, $\tilde{S}_{i,t}$
7: **end for**

Trovò et al. [56] presented theoretical guarantees both in the abruptly-changing environment, i.e., piece-wise stationary rewards with breakpoints, and in the smoothly-changing

environment, i.e., when the reward evolution can be associated to a continuous mathematical function (e.g., sinusoidal):

Theorem 3.3 (SW-TS Regret Abruptly [56]). *In an abruptly-changing MAB, with a number of breakpoints in the order of* $\mathcal{O}(T^{\alpha})$ *and* $\alpha \in [0, 1)$*, Algorithm 3.3 with window size* $h \propto T^{\frac{1-\alpha}{2}}$ *obtains a cumulative expected regret, dropping logarithmic terms:*

$$R_{\boldsymbol{\mu}}(\textit{SW-TS}, T) = \mathcal{O}\left(T^{\frac{1+\alpha}{2}}\right).$$

Theorem 3.4 (SW-TS Regret Smoothly [56]). *In a smoothly-changing MAB, where the difference between the rewards of two arms is smaller than* Δ *only for a limited number of rounds* $\mathcal{F}_{\Delta}$*, Algorithm 3.3 with window size* $h = T^{1-\beta}$ *obtains a cumulative expected regret, dropping logarithmic terms:*

$$R_{\boldsymbol{\mu}}(\textit{SW-TS}, T) = \mathcal{O}\left(T^{\beta}\right),$$

where $\beta \in [1 - \log_T\left(\frac{\Delta}{2\sigma}\right), 1]$.

Rexp3

An approach which does not rely on sliding-window filtering is the one propoesed by Besbes et al. [11]. The main idea behind their work is to exploit an efficient adversarial MAB algorithm, `Exp3` [7], introducing a series of epochs of size Δ_T in which the state of the algorithm is reset.

Algorithm 3.4 `Rexp3`

Input: T horizon, K arms, Δ_T epoch size, $\gamma \in [0,1]$.

Set batch index $j = 1$. // epochs index
while $j \leq \lceil T/\Delta_T \rceil$ **do**
  Set $\tau = (j-1)\Delta_T$.
  Initialization: $w_i^t = 1$ for each $i \in [K]$.
  // `Exp3` routine in epoch j
  for $t = \tau + 1, \ldots, \min\{T, \tau + \Delta_T\}$ **do**
    For each $i \in [K]$ set:

$$p_i^t = (1-\gamma)\frac{w_i^t}{\sum_{k\in[K]} w_k^t} + \frac{\gamma}{K}$$

    Draw an arm I_t from $[K]$ according to the distribution $\{p_i^t\}_{i\in[K]}$.
    Pull arm I_t and observe reward R_t.
    Set $\hat{X}_{I_t}^t = R_t/p_{I_t}^t$, and for any other $i \neq I_t$ set $\hat{X}_i^t = 0$.
    For all $i \in [K]$ update:

$$w_i^{t+1} = w_i^t \exp\left\{\frac{\gamma \hat{X}_i^t}{K}\right\}$$

  end for
  Set $j = j + 1$ // new epoch
end while

`Rexp3` is guaranteed to achieve a worst-case sublinear regret under the assumption that the total variation V_T of the arms' payoffs is known, i.e., $\sum_{t=2}^{T} \max_{i\in[K]} |\mu_i(t) - \mu_i(t-1)| \leq V_T$. This result was later improved by Chen et al. [17], who proved the regret sublinearity of `Rexp3` without requiring the knowledge of V_T.

Theorem 3.5 (`Rexp3` Regret [11]). *In a (non-stationary) stochastic MAB, the regret of Algorithm 3.4 with epoch size* $\Delta_T = \lceil (K \log K)^{1/3}(T/V_T)^{2/3} \rceil$ *and* $\gamma = \min\left\{1, \sqrt{\frac{K \log K}{(e-1)\Delta_T}}\right\}$ *satisfies:*

$$R_{\boldsymbol{\mu}}(\textit{Rexp3}, T) \leq \mathcal{O}\left((V_T K \log K)^{1/3} T^{2/3}\right),$$

for every $T \geq 1$, $K \geq 2$, $V_T \in [K^{-1}, K^{-1}T]$.

Ser4

Another approach, which consists of the idea of progressively reduce the arm set removing the ones that with high probability are not optimal, was the one proposed by Allesiardo

et al. [5], who developed the `Ser4` algorithm; at each round t, in order to deal with non-stationary processes and possibly best-arm switches, with a low probability ϕ, the set is restored to the original.

Algorithm 3.5 `Ser4`

Input: T horizon, K arms, $\delta \in [0,1]$, $\phi \in [0,1]$, $\epsilon \in \mathbb{R}^+$

Set $S_1 = [K]$, $\widehat{\mu}_i(0) = 0$ for each $i \in [K]$, $t = 1$, $\tau = 1$.
while $t \leq T$ **do**
  Shuffle S_τ. // set of active arms
  for all $i \in S_\tau$ **do**
    Pull $I_t = i$ and observe R_t.
    Update $\widehat{\mu}_i(\tau) = \frac{\tau-1}{\tau}\widehat{\mu}_i(\tau-1) + \frac{R_t}{\tau}$.
    $t = t + 1$.
  end for
  $i_{\max} = \text{argmax}_{i\in[K]}\, \widehat{\mu}_i(\tau)$.
  Remove from $S_{\tau+1}$ all i such that:

$$\widehat{\mu}_{i_{\max}}(\tau) - \widehat{\mu}_i + \epsilon \geq 2\sqrt{\frac{1}{2\tau}\log\left(\frac{4K\tau^2}{\delta}\right)}$$

  $\tau = \tau + 1$
  $t = t + 1$
  with probability ϕ
    $S_t = [K]$
    $\widehat{\mu}_i(t) = 0$ for each $i \in [K]$
    $\tau = 1$
  end with probability
end while

Theorem 3.6 (`Ser4` Regret [5]). *The expected cumulative regret of Algorithm 3.5 with* $\delta = 1/T$, $\tau_{min} = \log(KT)$, $\Delta \geq \frac{1}{KT}$, $\phi = \sqrt{\frac{N}{TK\log(KT)}}$ *and* $\epsilon \in [0,1)$ *is:*

$$R_{\boldsymbol{\mu}}(\texttt{Ser4}, T) \leq \min\left(\mathcal{O}\left(\frac{\sqrt{NTK\log(KT)}}{\Delta}\right), \mathcal{O}\left(T^{2/3}\sqrt{NK\log\frac{T}{K}}\right)\right),$$

where N is the number of best-arm switches.

3.2.2. Algorithms for Rested MABs

The *rested MAB* framework is less studied: the works which are more related to our setting are for sure the ones by Levine et al. [32] and Seznec et al. [51, 52] which discussed the *rotting* bandits case.

RAW-UCB

Seznec et al. [52] proposed an algorithm for rotting bandits in both the restless and the rested formulation. We present here the latter. The algorithm, called `RAW-UCB`, relies on the construction of an upper bound to the rewards, exploiting once again the optimistic approach (Section 1.3.2); since the rewards are decreasing, the sample mean $\widehat{\mu}_i(n)$ at pull n is already an upper bound of the payoff $\mu_i(n+1)$ and no additional exploration bonus is needed. Moreover, to update the sample means more accurately, `RAW-UCB` applies a sliding window h to the observations in order to consider for each arm only the most recent rewards, filtering out the outdated ones. The main difference between `RAW-UCB` and the previously introduced sliding window algorithms (`SW-UCB`, `SW-TS`, `SW-KL-UCB`) is that the size h is not fixed nor global, indeed it is determined at each step and for each arm in order to minimize the value of the upper bound for the given arm. An assumption required in order to guarantee the theoretical results on the regret of the algorithm is that the maximum payoff decay is bounded, i.e., for any arm i, any two consecutive pulls have a difference of the payoffs at most equal to a quantity L:

$$L = \max_{i\in[K]} \max_{n\in[T]} \{\mu_i(n+1) - \mu_i(n)\}.$$

Let us introduce the following quantities:

$$\widehat{\mu}_i^h(t) = \frac{1}{h} \sum_{l=1}^{t-1} R_l \mathbb{1}\{I_l = i \wedge N_{i,l} > N_{i,t-1} - h\}, \tag{3.8}$$

$$c(h, \delta_t) \coloneqq \sqrt{2\sigma^2 \log(2/\delta_t)/h}, \tag{3.9}$$

$$\mathrm{ind}(i, t, \delta_t) \coloneqq \min_{h \le N_{i,t-1}} \widehat{\mu}_i^h(t) + c(h, \delta_t), \tag{3.10}$$

Where Equation (3.8) represents the sample mean of the last h rewards obtained from arm i, Equation (3.9) is a confidence parameter to cope with stochasticity, and Equation (3.10) is the upper bound of arm i at round t.

Algorithm 3.6 RAW-UCB

Input: T horizon, K arms, $\sigma \in \mathbb{R}^+$, $\alpha \in \mathbb{N}$

for $t = 1, \dots, K$ **do**
  Pull arm $I_t = t$ and observe R_t.
  For each $h \in [N_{I_t,t}]$ update $\left\{\widehat{\mu}^h_{I_t}\right\}_h$ using Equation (3.8).
end for
for $t = K+1, \dots, T$ **do**
  Pull $I_t = \text{argmax}_{i \in [K]}\{\text{ind}(i, t, 2t^{-\alpha})\}$ and receive observation R_t.
  For each $h \in [N_{I_t,t}]$ update $\left\{\widehat{\mu}^h_{I_t}\right\}_h$ using Equation (3.8).
end for

Theorem 3.7 (RAW-UCB Regret [52]). *For any rotting bandit with non-increasing payoffs (Assumption 3.1) and bounded decay L, Algorithm 3.6 tuned with $\alpha \geq 5$ obtains a cumulative expected regret:*

$$R_{\boldsymbol{\mu}}(\textit{RAW-UCB}, T) \leq \mathcal{O}\left(KL + \sqrt{\sigma^2 \log T}\left(K + \sqrt{KT}\right)\right).$$

4 | Rising Bandits Analysis

In this chapter, we analyze both the *restless rising* bandit and the *rested rising* bandit settings, proposing a novel optimistic algorithm and discussing the theoretical results on the regret bound. For the seek of understanding, we present the *rested* setting first, being the *restless* a straightforward derivation of the former. Further details on the technical derivations and the full proofs of Theorems and Lemmas are available in Appendix A.

4.1. Rising Rested Bandits

This section is about the *Rising rested* bandits in which the payoff increases only when the arm is pulled, i.e., $\mu_i(t, N_{i,t-1}) = \mu_i(N_{i,t-1})$.

Oracle Policy First of all, we have to introduce the optimal policy for the rising rested setting. One can be tempted to use the *oracle greedy* policy, i.e., the policy which at each round t selects the arm with the highest payoff, however this is not an optimal choice for the rising rested setting. Intuitively, since the payoffs at a certain round depend on the previous pulls, a certain arm may have reached a higher payoff only because it were pulled more, hence comparing the payoffs at each round and selecting the highest one is pointless. We recall the *oracle constant* policy, i.e., the policy which at each round t selects the fixed arm that maximizes the sum of the payoffs over the horizon T, introduced by Heidari et al. [25] and proved to be optimal for non-decreasing rested bandits.

Theorem 4.1 (Constant Policy Optimality [25]). *Let $\pi^c_{\boldsymbol{\mu},T}$ be the* oracle constant *policy:*

$$\pi^c_{\boldsymbol{\mu},T}(t) \in \operatorname*{argmax}_{i\in[K]} \left\{ \sum_{l\in[T]} \mu_i(N_{i,l-1}) \right\}, \quad \forall t \in [T].$$

Then, $\pi^c_{\boldsymbol{\mu},T}$ is optimal for the rested non-decreasing bandits (i.e., under Assumption 2.1).

Theorem 4.1 establishes the first important difference between the non-decreasing and non-increasing (a.k.a. rotting) rested bandits. Indeed, while for the former ones the oracle

constant policy is optimal, for the latter, the oracle *greedy* policy is needed to achieve optimality [32].

Non-Learnability We now prove a result highlighting the "hardness" of non-decreasing rested bandits. We show that with no assumptions on the payoff $\mu_i(n)$ (e.g., concavity), it is not possible to devise a no-regret algorithm.

Theorem 4.2 (Non-Learnability). *There exists a 2-armed non-decreasing (non-concave) deterministic rested bandit with $\gamma_i(n) \leq \gamma_{\max} < 1$ for all $i \in [K]$ and $n \in \mathbb{N}$, such that any learning policy π suffers regret:*

$$R_{\boldsymbol{\mu}}(\pi, T) \geq \left\lfloor \frac{\gamma_{\max}}{12} T \right\rfloor .$$

The intuition behind this result is that, if we enforce no condition on how the increment $\gamma_i(n)$ will behave in the future by pulling the arm, we may never pull it to arrive to the point in which it becomes convenient to play it. Conversely, we might pull an arm for the sole purpose of making it evolve. Thus, Theorem 4.2 highlights the importance of the concavity assumption (Assumption 2.2) for the rested setting, providing an answer to an open question posed in Heidari et al. [25].

While being essential for the non-decreasing rested setting, Assumption 2.2 is less crucial for non-decreasing restless (Section 4.2) bandits [11].

4.1.1. Upper Bound Derivation[1]

Before introducing the algorithm and provide its regret analysis, we focus on the derivation of an upper bound suitable for our context. Creating such an upper bound means identifying a quantity $B_i(t) \geq \mu_i(t)$ for each $i \in [K]$, $t \in [T]$.

Let us consider the case in which the observed rewards are **deterministic** realizations of the arms ($\sigma = 0$), i.e., no stochastic distribution is considered. Along with the discussion of the derivation, we provide Figure 4.1 to highlight the general idea behind the concepts to come.

In the determinisitc situation the agent at each round observes the true payoff of the pulled arm I_t, i.e., $R_t = \mu_{I_t}(N_{I_t,t})$. We will try to build an estimator of $\mu_i(t)$, namely $\overline{\mu}_i^{\texttt{R-ed}}(t)$, having observed the exact payoffs of arm i up to round $t-1$: $\{\mu_i(n)\}_{n=0}^{N_{i,t-1}}$.

[1] Detailed proofs of this Section are in Appendix A.1.

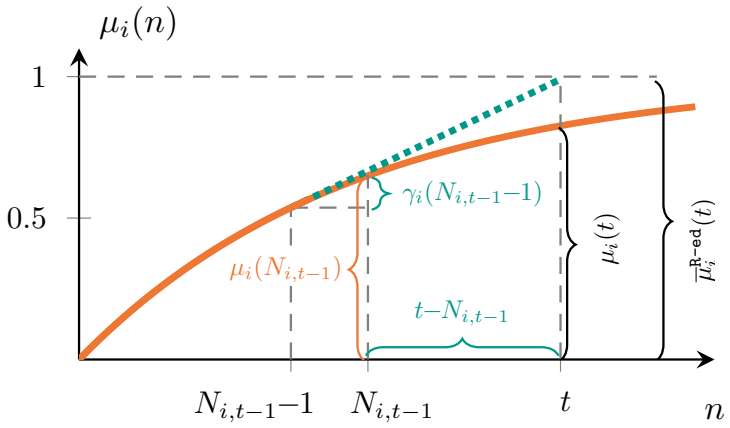


Figure 4.1: Upper Bound Construction: $\overline{\mu}_i^{\texttt{R-ed}}(t)$.

Assumption 2.1 (non-decreasing) allows deriving the following identity:

$$\mu_i(t) = \underbrace{\mu_i(N_{i,t-1})}_{\text{(most recent payoff)}} + \underbrace{\sum_{n=N_{i,t-1}}^{t-1} \gamma_i(n)}_{\text{(sum of future increments)}} .$$

Where the payoff of arm i at round t, $\mu_i(t)$, is expressed as the sum of the last observed payoff from arm i, $\mu_i(N_{i,t-1})$, and, the sum of its future increments $\gamma_i(n)$. It is worth noticing that $\mu_i(t)$ is the reward which would have been obtained if the agent pulled only arm i up to round t, i.e., $N_{i,t-1} = t-1$.

By exploiting the concavity (Assumption 2.2), we upper bound the sum of future increments with the last experienced increment $\gamma_i(N_{i,t-1}-1)$ that is projected for the future $t - N_{i,t-1}$ pulls:

$$\sum_{n=N_{i,t-1}}^{t-1} \gamma_i(n) \leq (t - N_{i,t-1})\gamma_i(N_{i,t-1}-1).$$

We have now obtained the upper bound for the *deterministic rested* rising bandit setting:

$$\overline{\mu}_i^{\texttt{R-ed}}(t) := \begin{cases} \underbrace{\mu_i(N_{i,t-1})}_{\text{(most recent payoff)}} + (t - N_{i,t-1}) \underbrace{\gamma_i(N_{i,t-1}-1)}_{\text{(most recent increment)}} & \text{if } N_{i,t-1} \geq 2 \\ +\infty & \text{otherwise} \end{cases} . \tag{4.1}$$

The optimism of $\overline{\mu}_i^{\texttt{R-ed}}$ and a bias bound are proved in the following Lemma:

Lemma 4.1 (Deterministic Rested UB). *For every arm $i \in [K]$ and every round $t \in [T]$:*

$$B_i(t) \equiv \overline{\mu}_i^{\texttt{R-ed}}(t) \geq \mu_i(t).$$

Moreover if $N_{i,t-1} \geq 2$ it holds that:

$$\overline{\mu}_i^{\texttt{R-ed}}(t) - \mu_i(N_{i,t}) \leq (t - N_{i,t-1})\gamma_i(N_{i,t-1} - 1).$$

We can now move to the **stochastic** scenario ($\sigma > 0$) in which the obtained rewards are sampled from stochastic distributions. The key idea to deal with the uncertainty in the observations is to average multiple samples of the estimator $\overline{\mu}_i^{\texttt{R-ed}}(t)$ in an h-wide window.

For every $l \in \{2, \dots, N_{i,t-1}\}$, we have that:

$$\mu_i(t) = \underbrace{\mu_i(l)}_{\text{(past payoff)}} + \underbrace{\sum_{j=l}^{t-1} \gamma_i(j)}_{\text{(sum of future increments)}} \leq \underbrace{\mu_i(l)}_{\text{(past payoff)}} + (t-l) \underbrace{\gamma_i(l-1)}_{\text{(past increment)}}$$

where the inequality follows from Assumption 2.2.[2]

Unfortunately, we do not have access to the exact payoffs $\mu_i(l)$ and exact increments $\gamma_i(l-1) = \mu_i(l) - \mu_i(l-1)$, but only to their corresponding point estimates $R_{t_{i,l}}$ and $R_{t_{i,l}} - R_{t_{i,l-1}}$, hence for a correct estimation we need to average over an h-wide window, replacing $\mu_i(l)$ and $\gamma_i(l-1)$ with quantities which provide suitable concentration properties. For $\mu_i(l)$ the substitution with $R_{t_{i,l}}$ is straightforward, while for $\gamma_i(l-1)$ the estimate $R_{i,t_l} - R_{i,t_{l-1}}$ is too unstable.

The following relation helps us solve this task and is especially important in the bandit derivations to come:

Lemma 4.2 (Growth Bound). *Under Assumptions 2.1 and 2.2, for every $i \in [K]$, $k, k' \in \mathbb{N}$ with $k' < k$, for both restless and rested bandits, it holds that:*

$$\gamma_i(k) \leq \frac{\mu_i(k) - \mu_i(k')}{k - k'}.$$

[2] The estimator of the deterministic case in Equation 4.1 is obtained by setting $l = N_{i,t-1}$.

Based on Lemma 4.2, we bound for every $l \in \{2, \dots, N_{i,t-1}\}$ and $h \in [l-1]$:

$$\underbrace{\gamma_i(l-1)}_{\text{(past increment at } l)} \leq \underbrace{\frac{\mu_i(l) - \mu_i(l-h)}{h}}_{\text{(average past increment over } \{l-h,\dots,l\})}.$$

We can now introduce $\widetilde{\mu}_i^{\texttt{R-ed},h}(t)$ and $\widehat{\mu}_i^{\texttt{R-ed},h}(t)$, the optimistic approximation of $\mu_i(t)$ and the corresponding estimator, defined when the window size is $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$:

$$\widetilde{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\underbrace{\mu_i(l)}_{\text{(past payoff)}} + (t-l) \underbrace{\frac{\mu_i(l) - \mu_i(l-h)}{h}}_{\text{(average past increment)}} \right),$$

$$\widehat{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\underbrace{R_{t_{i,l}}}_{\text{(estimated past payoff)}} + (t-l) \underbrace{\frac{R_{t_{i,l}} - R_{t_{i,l-h}}}{h}}_{\text{(estimated average past increment)}} \right).$$

An in-depth analysis of the above results leads to the following Lemmas:

- Lemma 4.3 shows that $\widetilde{\mu}_i^{\texttt{R-ed},h}(t)$ is an upper-bound for $\mu_i(t)$ and provides a bound to its bias for every value of h.
- Lemma 4.4 analyzes the concentration of $\widehat{\mu}_i^{\texttt{R-ed},h}(t)$ around $\widetilde{\mu}_i^{\texttt{R-ed},h}(t)$, for a specific choice of $\delta_t = t^{-\alpha}$ and when h is a function of the number of pulls $N_{i,t-1}$ only.

Lemma 4.3 (Stochastic Rested UB). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widetilde{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(l) + (t-l) \frac{\mu_i(l) - \mu_i(l-h)}{h} \right),$$

otherwise if $h = 0$, we set $\widetilde{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq +\infty$. Then, $\widetilde{\mu}_i^{\texttt{R-ed},h}(t) \geq \mu_i(t)$.

Moreover, if $N_{i,t-1} \geq 2$, it holds that:

$$\widetilde{\mu}_i^{\texttt{R-ed},h}(t) - \mu_i(N_{i,t}) \leq \frac{1}{2}(2t - 2N_{i,t-1} + h - 1)\gamma_i(N_{i,t-1} - 2h + 1).$$

Lemma 4.4 (Rested UB Concentration). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widehat{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(R_{t_{i,l}} + (t-l) \frac{R_{t_{i,l}} - R_{t_{i,l-h}}}{h} \right),$$

$$\beta_i^{\texttt{R-ed},h}(t,\delta) \coloneqq \sigma(t - N_{i,t-1} + h - 1) \sqrt{\frac{10 \log \frac{1}{\delta}}{h^3}},$$

otherwise if $h = 0$, we set $\widehat{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq +\infty$ and $\beta_i^{\texttt{R-ed},h}(t,\delta) \coloneqq +\infty$. 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:

$$\Pr\left(\left| \widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) \right| > \beta_i^{\texttt{R-ed},h_{i,t}}(t,\delta_t) \right) \leq 2t^{1-\alpha}.$$

4.1.2. Algorithm

We propose the following optimistic algorithm for Stochastic **R**ising Rest**ed** Bandits, called `R-ed-UCB`:

Algorithm 4.1 `R-ed-UCB`

Input: T horizon, K arms, $\sigma \in \mathbb{R}^+$,

1: Initialize $N_i \leftarrow 0$ for all $i \in [K]$
2: **for** $t = 1, \ldots, T$ **do**
3: Pull arm $I_t \in \text{argmax}_{i \in [K]} \{B_i(t)\}$.
4: Observe $R_t \sim \nu_{I_t}(N_{I_t,t-1})$.
5: Update B_{I_t} and $N_{I_t} \leftarrow N_{I_t} + 1$.
6: **end for**

where $B_i(t)$, called *exploration index*, is equivalent to the upper confidence bound presented in Section 4.1.1:

$$B_i(t) \equiv \begin{cases} \overline{\mu}_i^{\texttt{R-ed}}(t), & \text{if } \sigma = 0 \\ \widehat{\mu}_i^{\texttt{R-ed},h}(t) + \beta_i^{\texttt{R-ed},h}(t), & \text{otherwise} \end{cases}.$$

$f(l)$	e^{-cl}	l^{-c} $(cq > 1)$	l^{-c} $(cq = 1)$	l^{-c} $(cq \leq 1)$
$\mathcal{O}(\Upsilon_{\boldsymbol{\mu}}(M, q))$	$\frac{e^{-cq}}{cq}$	$\frac{1}{cq-1}$	$\log M$	$\frac{M^{1-cq}}{1-cq}$

Table 4.1: Big-O rates of $\Upsilon_{\boldsymbol{\mu}}(M, q)$ in the case $\gamma_i(l) \leq f(l)$ for all $i \in [K]$ and $l \in \mathbb{N}$ (see also Lemma 4.5).

4.1.3. Regret Analysis

In this section, we will provide an analysis of the cumulative regret obtained by `R-ed-UCB` both in the deterministic and in the stochastic setting, showing that the algorithm satisfies Property 1.4.

Problem Characterization To characterize the specific problem instance, we introduce the following quantity, called *cumulative increment*, defined for $q \in [0, 1]$ and $M \in [T]$ as:

$$\Upsilon_{\boldsymbol{\mu}}(M, q) \coloneqq \max_{i \in [K]} \left\{ \sum_{l=1}^{M-1} \gamma_i(l)^q \right\}. \tag{4.2}$$

The cumulative increment accounts for how fast the payoffs reach their asymptotic value, i.e., become stationary. Indeed, small values of $\Upsilon_{\boldsymbol{\mu}}(M, q)$ lead to simpler problems, as they are closer to stationary bandits. For particular choices of $\gamma_i(l)$ and q, it is possible to find some bounds on $\Upsilon_{\boldsymbol{\mu}}(T, q)$. Lemma 4.5 and Table 4.1 summarize some Big-O rates for $\Upsilon_{\boldsymbol{\mu}}(T, q)$ under certain choices of the growth rates γ_i.

Lemma 4.5. *Let $\Upsilon_{\boldsymbol{\mu}}(M, q)$ be as defined in Equation (4.2) for some $q \in [0, 1]$. Then, for all $i \in [K]$ and $l \in \mathbb{N}$ the following statements hold:*

- *if $\gamma_i(l) \leq be^{-cl}$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}\left(b^q \frac{e^{-cq}}{cq}\right)$;*
- *if $\gamma_i(l) \leq bl^{-c}$ with $cq > 1$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}\left(\frac{b^q}{cq-1}\right)$;*
- *if $\gamma_i(l) \leq bl^{-c}$ with $cq = 1$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}(b^q \log M)$;*
- *if $\gamma_i(l) \leq bl^{-c}$ with $cq < 1$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}\left(b^q \frac{M^{1-cq}}{1-cq}\right)$.*

Both in the rested and in the restless setting, the cumulative regret $R_{\boldsymbol{\mu}}$ can be bounded by a function of the horizon T, the number of arms K and the cumulative increment $\Upsilon_{\boldsymbol{\mu}}(M, q)$, hence, it is possible to tighten the bound by manipulating the value of q.

Deterministic Setting

We provide the analysis of R-ed-UCB when we employ the exploration index:

$$B_i(t) \equiv \overline{\mu}_i^{\texttt{R-ed}}(t).$$

Theorem 4.3. *Let $T \in \mathbb{N}$, then R-ed-UCB (Algorithm 4.1) with $B_i(t) \equiv \overline{\mu}_i^{\texttt{R-ed}}(t)$ suffers an expected regret bounded, for every $q \in [0, 1]$, as:*

$$R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T) \leq 2K + KT^q \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil \frac{T}{K} \right\rceil, q \right).$$

Proof. We have to analyze the following expression:

$$R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T) = \sum_{t=1}^{T} \mu_{i^*}(t) - \mu_{I_t}(N_{i,t}),$$

where $i^* \in \text{argmax}_{i \in [K]} \left\{ \sum_{l \in [T]} \mu_i(l) \right\}$.

We consider a term at a time, use $B_i(t) \equiv \overline{\mu}_i^{\texttt{R-ed}}(t)$, and we exploit the optimism, i.e., $B_{i^*}(t) \leq B_{I_t}(t)$:

$$\begin{aligned} \mu_{i^*}(t) - \mu_{I_t}(N_{I_t,t}) + B_{I_t}(t) - B_{I_t}(t) &\leq \min \left\{ 1, \underbrace{\mu_{i^*}(t) - B_{i^*}(t)}_{\leq 0} + B_{I_t}(t) - \mu_{I_t}(N_{I_t,t}) \right\} \\ &\leq \min \left\{ 1, B_{I_t}(t) - \mu_{I_t}(N_{I_t,t}) \right\}. \end{aligned}$$

Now we work on the term inside the minimum when $N_{I_t,t-1} \geq 2$:

$$B_{I_t}(t) - \mu_{I_t}(N_{I_t,t}) = \overline{\mu}_i^{\texttt{R-ed}}(t) - \mu_{I_t}(N_{I_t,t}) \leq (t - N_{i,t-1}) \gamma_{I_t}(N_{i,t-1} - 1),$$

where the inequality follows from Lemma 4.1.

We are going to decompose the summation of this term over the K arms:

$$\begin{aligned} R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T) &\leq \sum_{t=1}^{T} \min \left\{ 1, (t - N_{i,t-1}) \gamma_{I_t}(N_{i,t-1} - 1) \right\} \\ &\leq 2K + \sum_{i \in [K]} \sum_{j=3}^{N_{i,T}} \min \left\{ 1, (t_{i,j} - (j-1)) \gamma_i(j-2) \right\}, \end{aligned}$$

where $t_{i,j} \in [T]$ is the round at which arm $i \in [K]$ was pulled for the j-th time. Now, $q \in [0,1]$, then for any $x \geq 0$ it holds that $\min\{1,x\} \leq \min\{1,x\}^q \leq x^q$. By applying this latter inequality to the inner summation, we get:

$$\sum_{j=3}^{N_{i,T}} \min\left\{1, (t_{i,j} - (j-1))\gamma_i(j-2)\right\} \leq \sum_{j=3}^{N_{i,T}} \min\left\{1, T\gamma_i(j-2)\right\} \leq T^q \sum_{j=3}^{N_{i,T}} \gamma_i(j-2)^q,$$

having used $t_{i,j} - (j-1) \leq T$. Summing over the arms, we obtain:

$$T^q \sum_{i\in[K]} \sum_{j=3}^{N_{i,T}} \gamma_i(j-2)^q \leq T^q \sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}(N_{i,T}, q) \leq T^q K \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil \frac{T}{K} \right\rceil, q\right),$$

where the last inequality is obtained from Lemma A.3. □

Stochastic Setting

We provide the regret analysis of `R-ed-UCB` when we employ the exploration index:

$$B_i(t) \equiv \widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) + \beta_i^{\texttt{R-ed},h_{i,t}}(t),$$

where $\beta_i^{\texttt{R-ed},h}(t)$ is the exploration bonus defined in Lemma 4.4 and $h_{i,t}$ is an arm and time-dependent window. The following result provides the regret bound, under particular choices of $h_{i,t}$ and δ_t.

Theorem 4.4. *Let $T \in \mathbb{N}$, then `R-ed-UCB`(Algorithm 4.1) with $B_i(t) \equiv \widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) + \beta_i^{\texttt{R-ed},h_{i,t}}(t)$, $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$ and $\delta_t = t^{-\alpha}$ for $\alpha > 2$, suffers an expected regret bounded, for every $q \in [0,1]$, as:*

$$R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T) \leq \mathcal{O}\left(\frac{K}{\epsilon}(\sigma T)^{\frac{2}{3}} (\alpha \log T)^{\frac{1}{3}} + \frac{KT^q}{1-2\epsilon} \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q\right)\right).$$

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:

$$\mathcal{E}_{i,t} \coloneqq \left\{\left|\widetilde{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) - \widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t)\right| \leq \beta_i^{\texttt{R-ed},h_{i,t}}(t)\right\} \qquad \forall i \in [T],\, i \in [K].$$

We have to analyze the following expression:

$$R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i^*}(t) - \mu_{I_t}(N_{i,t})\right],$$

where $i^* \in \text{argmax}_{i\in[K]} \left\{\sum_{l\in[T]} \mu_i(l)\right\}$.

We decompose the above expression according to the good events $\mathcal{E}_t$:

$$\begin{aligned} R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T) &= \sum_{t=1}^{T} \mathbb{E}\left[(\mu_{i^*}(t) - \mu_{I_t}(N_{I_t,t}))\, \mathbb{1}\{\mathcal{E}_t\}\right] + \sum_{t=1}^{T} \mathbb{E}\left[(\mu_{i^*}(t) - \mu_{I_t}(N_{I_t,t}))\, \mathbb{1}\{\neg\mathcal{E}_t\}\right] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\left[(\mu_{i^*}(t) - \mu_{I_t}(N_{I_t,t}))\, \mathbb{1}\{\mathcal{E}_t\}\right] + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\}\right], \end{aligned}$$

where in the last line we exploited $\mu_{i^*}(t) - \mu_{I_t}(N_{I_t,t}) \leq 1$.

Now, we bound the second summation, recalling that $\alpha > 2$:

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\}\right] &= \sum_{t=1}^{T} \Pr\left(\neg\mathcal{E}_t\right) \\ &= 1 + \sum_{t=2}^{T} \Pr\left(\neg \bigcap_{i\in[K]} \mathcal{E}_{i,t}\right) \\ &= 1 + \sum_{t=2}^{T} \Pr\left(\bigcup_{i\in[K]} \neg\mathcal{E}_{i,t}\right) \\ &\leq 1 + \sum_{i\in[K]} \sum_{t=2}^{T} \Pr\left(\neg\mathcal{E}_{i,t}\right), \end{aligned}$$

where the first inequality is obtained with $\Pr(\neg\mathcal{E}_1) \leq 1$ and the second with a union bound over $[K]$.

Recalling $\Pr(\neg\mathcal{E}_{i,t})$ was bounded in Lemma 4.4, we bound the summation with the integral as in Lemma A.4 to get:

$$\sum_{i\in[K]} \sum_{t=2}^{T} \Pr\left(\neg\mathcal{E}_{i,t}\right) \leq \sum_{i\in[K]} \sum_{t=2}^{T} 2t^{1-\alpha} \leq 2K \int_{x=1}^{+\infty} x^{1-\alpha} \mathrm{d}x = \frac{2K}{\alpha - 2}.$$

From now on, we proceed the analysis under the good events $\mathcal{E}_t$, recalling that $B_i(t) \equiv \widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) + \beta_i^{\texttt{R-ed},h_{i,t}}(t, \delta_t)$. We consider each addendum of the summation and we exploit

the optimism, i.e., $B_{i^*}(t) \leq B_{I_t}(t)$:

$$\begin{aligned}\mu_{i^*}(t) - \mu_{I_t}(N_{I_t,t}) + B_{I_t}(t) - B_{I_t}(t) &\leq \min\left\{1, \underbrace{\mu_{i^*}(t) - B_{i^*}(t)}_{\leq 0} + B_{I_t}(t) - \mu_{I_t}(N_{I_t,t})\right\} \\ &\leq \min\left\{1, B_{I_t}(t) - \mu_{I_t}(N_{I_t,t})\right\}.\end{aligned}$$

Now, we work on the term inside the minimum:

$$\begin{aligned}B_{I_t}(t) - \mu_{I_t}(N_{I_t,t}) &= \widehat{\mu}_{I_t}^{\texttt{R-ed},h_{I_t,t}}(t) + \beta_{I_t}^{\texttt{R-ed},h_{I_t,t}}(t,\delta_t) - \mu_{I_t}(N_{I_t,t}) && (4.3)\\ &\leq \underbrace{\widetilde{\mu}_{I_t}^{\texttt{R-ed},h_{I_t,t}}(t) - \mu_{I_t}(N_{I_t,t})}_{\text{(a)}} + \underbrace{2\beta_{I_t}^{\texttt{R-ed},h_{I_t,t}}(t,\delta_t)}_{\text{(b)}}, && (4.4)\end{aligned}$$

where line (4.3) follows from the definition of $B_i(t)$, and line (4.4) derives from the fact that we are under the good event $\mathcal{E}_t$.

We now decompose over the arms and consider one term at a time, starting with (a):

$$\begin{aligned}&\sum_{t=1}^{T} \min\left\{1, \widetilde{\mu}_{I_t}^{\texttt{R-ed},h_{I_t,t}}(t) - \mu_{I_t}(N_{I_t,t})\right\} && (4.5)\\ &\quad\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_{i}^{\texttt{R-ed},h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(j)\right\} \\ &\quad\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{1}{2}(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)\gamma_i((j-1) - 2h_{i,t_{i,j}} + 1)\right\} && (4.6)\\ &\quad\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\left\{1, T\gamma_i(j - 2\lfloor\epsilon(j-1)\rfloor)\right\} && (4.7)\\ &\quad\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\left\{1, T\gamma_i(\lfloor(1-2\epsilon)j\rfloor)\right\} && (4.8)\\ &\quad\leq 2K + T^q \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \gamma_i(\lfloor(1-2\epsilon)j\rfloor)^q && (4.9)\\ &\quad\leq 2K + T^q \left\lceil\frac{1}{1-2\epsilon}\right\rceil \sum_{i\in[K]}\sum_{j=\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor} \gamma_i(j) && (4.10)\\ &\quad\leq 2K + T^q \left\lceil\frac{1}{1-2\epsilon}\right\rceil \sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(\lfloor(1-2\epsilon)N_{i,T}\rfloor, q\right) && (4.11)\\ &\quad\leq 2K + KT^q \left\lceil\frac{1}{1-2\epsilon}\right\rceil \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil, q\right), && (4.12)\end{aligned}$$

where line (4.6) follows from Lemma 4.3, line (4.7) is obtained by bounding $2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1 \leq 2T$ and exploiting the definition of $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$, line (4.8) follows from the observation $j - 2\lfloor \epsilon(j-1) \rfloor \geq j - 2\epsilon(j-1) \geq \lfloor (1-2\epsilon)j \rfloor$, line (4.9) is obtained from the already exploited inequality $\min\{1, x\} \leq \min\{1, x\}^q \leq x^q$ for $q \in [0, 1]$, line (4.10) is an application of Lemma A.2, line (4.11) applies the definition of $\Upsilon_{\boldsymbol{\mu}}(\cdot, q)$, and line (4.12) follows from Lemma A.3 recalling that $\sum_{i \in [K]} \lfloor (1-2\epsilon) N_{i,T} \rfloor \leq (1-2\epsilon)T$.

Let us now move to the concentration term (b). We decompose over the arms as well, taking care of the pulls in which $h_{i,j} = 0$, that are at most $1 + \left\lceil \frac{1}{\epsilon} \right\rceil$:

$$
\begin{aligned}
&\sum_{t=1}^{T} \min \left\{ 1, 2\beta_{I_t}^{\texttt{R-ed}, h_{I_t,t}}(t, \delta_t) \right\} \\
&\quad \leq K + K \left\lceil \frac{1}{\epsilon} \right\rceil + \sum_{i \in [K]} \sum_{j = \left\lceil \frac{1}{\epsilon} \right\rceil + 1}^{N_{i,T}} \min \left\{ 1, 2\sigma (t_{i,t} - (j-1) + h_{i,t_{i,t}} - 1) \sqrt{\frac{10 \log(t^\alpha)}{h_{i,t_{i,t}}^3}} \right\} \\
&\quad \leq K + K \left\lceil \frac{1}{\epsilon} \right\rceil + \sum_{i \in [K]} \sum_{j = \left\lceil \frac{1}{\epsilon} \right\rceil + 1}^{N_{i,T}} \min \left\{ 1, 2\sigma T \sqrt{\frac{10\alpha \log(T)}{\lfloor \epsilon(j-1) \rfloor^3}} \right\}, \qquad (4.13)
\end{aligned}
$$

where line (4.13) follows from bounding $t^\alpha \leq T^\alpha$ and from the definition of $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$.

To bound the summation, we compute the minimum integer value j^* (that turns out to be independent of i) of j such that the minimum is attained by its second argument:

$$
\begin{aligned}
2\sigma T \sqrt{\frac{10\alpha \log(T)}{\lfloor \epsilon(j-1) \rfloor^3}} \leq 1 \implies \lfloor \epsilon(j-1) \rfloor &\geq (2\sigma T)^{\frac{2}{3}} (10\alpha \log T)^{\frac{1}{3}} \\
\implies j^* &= \left\lceil \frac{1 + \epsilon + (2\sigma T)^{\frac{2}{3}} (10\alpha \log T)^{\frac{1}{3}}}{\epsilon} \right\rceil.
\end{aligned}
$$

Thus, we have:

$$
\begin{aligned}
&K + K\left\lceil \frac{1}{\epsilon} \right\rceil + \sum_{i\in[K]} \sum_{j=\left\lceil \frac{1}{\epsilon} \right\rceil + 1}^{N_{i,T}} \min\left\{1, 2\sigma T \sqrt{\frac{10\alpha \log(T)}{\lfloor \epsilon(j-1) \rfloor^3}}\right\} \\
&\leq K + K\left\lceil \frac{1}{\epsilon} \right\rceil + \sum_{i\in[K]} \left(\sum_{j=\left\lceil \frac{1}{\epsilon} \right\rceil + 1}^{j^*} 1 + \sum_{j=j^*+1}^{N_{i,T}} 2\sigma T \sqrt{\frac{10\alpha \log(T)}{\lfloor \epsilon(j-1) \rfloor^3}} \right) && (4.14) \\
&\leq K + K\left\lceil \frac{1}{\epsilon} \right\rceil + K\left(j^* - 1 - \left\lceil \frac{1}{\epsilon} \right\rceil + 1\right) + 2K\sigma T \sqrt{10\alpha \log(T)} \int_{x=j^*}^{+\infty} \frac{1}{(\epsilon(x-1)-1)^{\frac{3}{2}}} \mathrm{d}x && (4.15) \\
&= K + Kj^* + \frac{4K\sigma T \sqrt{10\alpha \log(T)}}{\epsilon\left(\epsilon(j^*-1)-1\right)^{\frac{1}{2}}} \\
&= K\left(3 + \frac{1}{\epsilon}\right) + \frac{3K}{\epsilon}(2\sigma T)^{\frac{2}{3}} \left(10\alpha \log T\right)^{\frac{1}{3}}, && (4.16)
\end{aligned}
$$

where line (4.14) is obtained by splitting the summation based on the value of j^*, line (4.15) comes from bounding the summation with the integral (Lemma A.4), and line (4.16) follows from substituting the value of j^* and simple algebraic manipulations.

Putting all together, we obtain:

$$
\begin{aligned}
R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T) &\leq 1 + \frac{2K}{\alpha - 2} + 5K + \frac{K}{\epsilon} + \frac{3K}{\epsilon}(2\sigma T)^{\frac{2}{3}} \left(10\alpha \log T\right)^{\frac{1}{3}} \\
&\quad + KT^q \left\lceil \frac{1}{1-2\epsilon} \right\rceil \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q\right) \\
&= \mathcal{O}\left(\frac{K}{\epsilon}(\sigma T)^{\frac{2}{3}} \left(\alpha \log T\right)^{\frac{1}{3}} + \frac{KT^q}{1-2\epsilon} \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q\right)\right).
\end{aligned}
$$

□

4.2. Rising Restless Bandits

This section is about the *Rising restless* bandits in which the payoff increases at every round regardless the arm is pulled, i.e., $\mu_i(t, N_{i,t-1}) = \mu_i(t)$.

Oracle Policy First of all, we have to introduce the optimal policy for the rising restless setting: the *oracle greedy* policy, i.e., the policy which at each round $t \in [T]$ selects the arm with the largest payoff in that round, is optimal for the non-decreasing restless bandit setting.

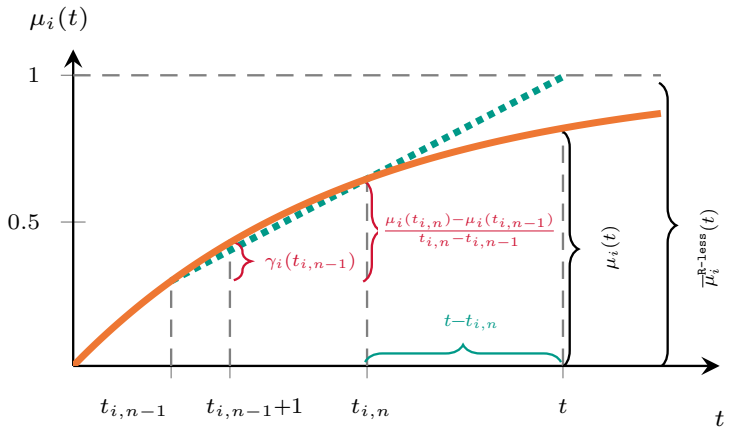


Figure 4.2: Upper Bound Construction: $\overline{\mu}_i^{\texttt{R-less}}(t)$.

Theorem 4.5 (Greedy Policy Optimality [52]). *Let $\pi^g_{\boldsymbol{\mu}}$ be the* oracle greedy *policy:*

$$\pi^g_{\boldsymbol{\mu}}(t) \in \underset{i\in[K]}{\text{argmax}}\{\mu_i(t)\}, \quad \forall t \in [T].$$

Then, $\pi^g_{\boldsymbol{\mu}}$ is optimal for the restless non-decreasing bandits (i.e., under Assumption 2.1).

It is worth noticing that the concavity assumption (Assumption 2.2) is not required, indeed the non-decreasing payoff assumption (Assumption 2.1) is enough to ensure the optimality of the greedy policy $\pi^g_{\boldsymbol{\mu}}$. This policy is also known to be optimal for restless and rested *rotting* bandits [32], but, as we already saw in Section 4.1, is not optimal in the rested *rising* setting.

4.2.1. Upper Bound Derivation[3]

Once again, before introducing the algorithm and provide its regret analysis, we focus on the derivation of an upper bound suitable for the restless context.

Let's start with the case in which the observed rewards are **deterministic** realizations of the arms ($\sigma = 0$), i.e., no stochastic distribution is considered. In this situation the agent at each round observes the true payoff of the pulled arm I_t, i.e., $R_t = \mu_{I_t}(t)$. Figure 4.2 highlights the general idea behind the concepts to come.

From now on, we will make use of the notation $t_{i,n}$ to refer to the round in which arm i was pulled for the n-th time.

Similarly to the rested case, we design $\overline{\mu}_i^{\texttt{R-less}}(t)$, an optimistic estimator of $\mu_i(t)$, employ-

[3]Detailed proofs of this Section are in Appendix A.2.

ing the exact payoffs observed up to round $t-1$: $\{\mu_i(t_{i,n})\}_{n=1}^{N_{i,t-1}}$.

We exploit the non-decreasing assumption (Assumption 2.1) to derive the identity:

$$\mu_i(t) = \underbrace{\mu_i(t_{i,N_{i,t-1}})}_{\text{(most recent payoff)}} + \underbrace{\sum_{l=t_{i,N_{i,t-1}}}^{t-1} \gamma_i(l)}_{\text{(sum of future increments)}} .$$

Then, we use the concavity (Assumption 2.2) to upper bound the sum of future increments with the last experienced increment that will be projected in the future for $t - t_{i,N_{i,t-1}}$ rounds:

$$\begin{aligned}\sum_{l=t_{i,N_{i,t-1}}}^{t-1} \gamma_i(l) &\leq (t - t_{i,N_{i,t-1}})\gamma_i(t_{i,N_{i,t-1}}) \\ &\leq (t - t_{i,N_{i,t-1}})\frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}},\end{aligned}$$

where the first inequality is obtained from Assumption 2.2, and the last inequality from Lemma 4.2.

Hence, the final estimator is:

$$\overline{\mu}_i^{\texttt{R-less}}(t) := \begin{cases} \underbrace{\mu_i(t_{i,N_{i,t-1}})}_{\text{(most recent payoff)}} + (t - t_{i,N_{i,t-1}}) \underbrace{\frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}}}_{\text{(most recent increment)}} & \text{if } N_{i,t-1} \geq 2, \\ +\infty & \text{otherwise.} \end{cases} .$$

Lemma 4.6 shows that $\overline{\mu}_i^{\texttt{R-less}}$ is optimistic and provides a bias bound.

Lemma 4.6 (Deterministic Restless UB). *For every arm $i \in [K]$ and every round $t \in [T]$:*

$$B_i(t) \equiv \overline{\mu}_i^{\texttt{R-less}}(t) \geq \mu_i(t).$$

Moreover, if $N_{i,t-1} \geq 2$, it holds that:

$$\overline{\mu}_i^{\texttt{R-less}}(t) - \mu_i(t) \leq \left(t - t_{i,N_{i,t-1}}\right) \gamma_i(t_{i,N_{i,t-1}-1}).$$

We can now move to the **stochastic** scenario ($\sigma > 0$) in which the reward obtained by the agent are sampled from a stochastic distribution. Here we can no longer exploit the

estimator $\overline{\mu}_i^{\texttt{R-less}}(t)$ since we only observe the noisy versions of μ_i, i.e., $\{R_{t_{i,n}}\}_{n=1}^{N_{i,t-1}}$. To cope with stochasticity we introduce an h-wide window to consider only the h most recent samples.

For every $l \in \{2, \dots, N_{i,t-1}\}$, we have that:

$$\mu_i(t) = \underbrace{\mu_i(t_{i,l})}_{\text{(past payoff)}} + \underbrace{\sum_{j=t_l}^{t-1} \gamma_i(j)}_{\text{(sum of future increments)}} \leq \underbrace{\mu_i(l)}_{\text{(past payoff)}} + (t - t_{i,l}) \underbrace{\gamma_i(t_{i,l-1})}_{\text{(past increment)}} .$$

where the inequality follows from Assumption 2.2.

Differently from the rested setting, we may not have direct access to the "instantaneous" growth $\gamma_i(t_{i,l-1})$ since the arm may have not been pulled two consecutive times, hence, we need to perform a further bounding step. Specifically, based on Lemma 4.2, we bound for every $l \in \{2, \dots, N_{i,t-1}\}$ and $h \in [l-1]$:

$$\underbrace{\gamma_i(t_{i,l-1})}_{\text{(past increment at } t_{i,l})} \leq \underbrace{\frac{\mu_i(t_{i,l}) - \mu_i(t_{i,l-h})}{t_{i,l} - t_{i,l-h}}}_{\text{(average past increment over } \{t_{i,l-h}, \dots, t_{i,l}\})} .$$

We report $\widetilde{\widetilde{\mu}}_i^{\texttt{R-ed},h}(t)$ and $\widehat{\widehat{\mu}}_i^{\texttt{R-ed},h}(t)$, a first proposal for the optimistic approximation of $\mu_i(t)$ and the corresponding estimator, defined when the window is of size $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$:

$$\widetilde{\widetilde{\mu}}_i^{\texttt{R-less},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\underbrace{\mu_i(t_{i,l})}_{\text{(past payoff)}} + (t - t_{i,l}) \underbrace{\frac{\mu_i(t_{i,l}) - \mu_i(t_{i,l-h})}{t_{i,l} - t_{i,l-h}}}_{\text{(average past increment)}} \right),$$

$$\widehat{\widehat{\mu}}_i^{\texttt{R-less},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\underbrace{R_{t_{i,l}}}_{\text{(estimated past payoff)}} + (t - t_{i,l}) \underbrace{\frac{R_{t_{i,l}} - R_{t_{i,l-h}}}{t_{i,l} - t_{i,l-h}}}_{\text{(estimated average past increment)}} \right).$$

Unfortunately $\widehat{\widehat{\mu}}_i^{\texttt{R-ed},h}(t)$, although intuitive, does not enjoy desirable concentration properties due to the presence of the denominator $t_{i,l} - t_{i,l-h}$ that is inconveniently correlated with the numerator $R_{t_{i,l}} - R_{t_{i,l-h}}$. For this reason, we resort to different estimators, with

better concentration properties but larger bias:

$$\widetilde{\mu}_i^{\texttt{R-less},h}(t) \coloneqq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\underbrace{\mu_i(t_{i,l})}_{\text{(past payoff)}} + (t-l)\underbrace{\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h}}_{\text{(average past increment)}}\right),$$

$$\widehat{\mu}_i^{\texttt{R-less},h}(t) \coloneqq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\underbrace{R_{t_{i,l}}}_{\text{(estimated past payoff)}} + (t-l)\underbrace{\frac{R_{t_{i,l}}-R_{t_{i,l-h}}}{h}}_{\text{(estimated average past increment)}}\right).$$

These estimators are actually upper-bounds of the previous ones since $t_{i,l} - t_{i,l-h} \geq h$ and $t_{i,l} \geq l$.

An in-depth analysis of the above results leads to the following Lemmas:

- Lemma 4.7 shows that $\widetilde{\mu}_i^{\texttt{R-less},h}(t)$ is an upper-bound for $\mu_i(t)$ and provides a bound to its bias for every value of h;
- Lemma 4.8 analyzes the concentration of $\widehat{\mu}_i^{\texttt{R-less},h}(t)$ around $\widetilde{\mu}_i^{\texttt{R-less},h}(t)$ for a specific choice of $\delta_t = t^{-\alpha}$ and when h is a function of the number of pulls $N_{i,t-1}$ only;

Lemma 4.7 (Stochastic Restless UB). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widetilde{\mu}_i^{\texttt{R-less},h}(t) :- \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l}) + (t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h}\right),$$

otherwise if $h = 0$, we set $\widetilde{\mu}_i^{\texttt{R-less},h}(t) \coloneqq +\infty$. Then, $\widetilde{\mu}_i^{\texttt{R-less},h}(t) \geq \mu_i(t_{i,N_{i,t-1}})$.

Moreover, if $N_{i,t-1} \geq 2$ it holds that:

$$\widetilde{\mu}_i^{\texttt{R-less},h}(t) - \mu_i(t) \leq \frac{(2t - 2N_{i,t-1} + h - 1)(t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-2h+1})}{2h}\gamma_i(t_{i,N_{i,t-1}-2h+1}).$$

Lemma 4.8 (Restless UB Concentration). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widehat{\mu}_i^{\texttt{R-less},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(R_{t_{i,l}} + (t-l) \frac{R_{t_{i,l}} - R_{t_{i,l-h}}}{h} \right),$$

$$\beta_i^{\texttt{R-less},h}(t,\delta) \coloneqq \sigma(t - N_{i,t-1} + h - 1) \sqrt{\frac{10 \log \frac{1}{\delta}}{h^3}},$$

otherwise if $h = 0$, we set $\widehat{\mu}_i^{\texttt{R-less},h}(t) \coloneqq +\infty$ and $\beta_i^{\texttt{R-less},h}(t,\delta) \coloneqq +\infty$. 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:

$$\Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) - \widetilde{\mu}_i^{\texttt{R-less},h_{i,t}}(t) \right| > \beta_i^{\texttt{R-less},h_{i,t}}(t,\delta_t) \right) \leq 2t^{1-\alpha}.$$

4.2.2. Algorithm

We propose the following optimistic algorithm for Stochastic **R**ising Rest**less** Bandits, called `R-less-UCB`:

Algorithm 4.2 `R-less-UCB`

Input: T horizon, K arms, $\sigma \in \mathbb{R}^+$,

1: Initialize $N_i \leftarrow 0$ for all $i \in [K]$
2: **for** $t = 1, \dots, T$ **do**
3: Pull arm $I_t \in \operatorname{argmax}_{i \in [K]} \{B_i(t)\}$.
4: Observe $R_t \sim \nu_{I_t}(t)$.
5: Update B_{I_t} and $N_{I_t} \leftarrow N_{I_t} + 1$.
6: **end for**

where $B_i(t)$, the *exploration index*, is equivalent to the upper confidence bound presented in Section 4.2.1:

$$B_i(t) \equiv \begin{cases} \overline{\mu}_i^{\texttt{R-less}}(t), & \text{if } \sigma = 0 \\ \widehat{\mu}_i^{\texttt{R-less},h}(t) + \beta_i^{\texttt{R-less},h}(t), & \text{otherwise} \end{cases}.$$

4.2.3. Regret Analysis

In this Section we will provide an analysis of the cumulative expected regret obtained by `R-less-UCB` both in the deterministic and in the stochastic setting, showing that the algorithm satisfies Property 1.4.

Deterministic Setting

We provide the regret analysis of `R-less-UCB` when we employ the exploration index:

$$B_i(t) \equiv \overline{\mu}_i^{\texttt{R-less}}(t).$$

Theorem 4.6. *Let $T \in \mathbb{N}$, then `R-less-UCB`(Algorithm 4.2) with $B_i(t) \equiv \overline{\mu}_i^{\texttt{R-less}}(t)$ suffers an expected regret bounded, for every $q \in [0,1]$, as:*

$$R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) \leq 2K + KT^{\frac{q}{q+1}} \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil \frac{T}{K} \right\rceil, q \right)^{\frac{1}{q+1}}.$$

Proof. We have to analyze the following expression:

$$R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) = \sum_{t=1}^{T} \mu_{i_t^*}(t) - \mu_{I_t}(t),$$

where $i_t^* \in \operatorname{argmax}_{i \in [K]} \{\mu_i(t)\}$ for all $t \in [T]$.

We consider each round at a time, recalling that $B_i(t) \equiv \overline{\mu}_i^{\texttt{R-less}}(t)$, and using optimism, i.e., $B_{i_t^*}(t) \leq B_{I_t}(t)$, we have:

$$\begin{aligned} \mu_{i^*}(t) - \mu_{I_t}(t) + B_{I_t}(t) - B_{I_t}(t) &\leq \min \left\{ 1, \underbrace{\mu_{i_t^*}(t) - B_{i_t^*}(t)}_{\leq 0} + B_{I_t}(t) - \mu_{I_t}(t) \right\} \\ &\leq \min \left\{ 1, B_{I_t}(t) - \mu_{I_t}(t) \right\}. \end{aligned} \tag{4.17}$$

Now we consider the term inside the minimum, when $N_{I_t, t-1} \geq 2$:

$$\begin{aligned} B_{I_t}(t) - \mu_{I_t}(t) &= \overline{\mu}_{I_t}^{\texttt{R-less}}(t) - \mu_{I_t}(t) && (4.18) \\ &\leq (t - t_{i, N_{i,t-1}}) \gamma_i(t_{i, N_{i,t-1}-1}), && (4.19) \end{aligned}$$

where to get line (4.19) we applied Lemma 4.6.

Let us plug the expression derived in Equation (4.17) and decompose the summation of this term w.r.t. the K arms:

$$
\begin{aligned}
R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) &\leq \sum_{t=1}^{T} \min\left\{1, (t - t_{i,N_{i,t-1}})\gamma_i(t_{i,N_{i,t-1}-1}), \right\} \\
&= 2K + \sum_{i\in[K]} \sum_{j=3}^{N_{i,T}} \min\left\{1, (t_{i,j} - t_{i,j-1})\gamma_i(t_{i,j-2})\right\} \\
&\leq 2K + \sum_{i\in[K]} \sum_{j=3}^{N_{i,T}} (t_{i,j} - t_{i,j-1})^y \gamma_i(t_{i,j-2})^y && (4.20) \\
&\leq 2K + \sum_{i\in[K]} \left(\sum_{j=3}^{N_{i,T}} (t_{i,j} - t_{i,j-1})\right)^y \left(\sum_{j=3}^{N_{i,T}} \gamma_i(t_{i,j-2})^{\frac{y}{1-y}}\right)^{1-y} && (4.21) \\
&\leq 2K + T^y \sum_{i\in[K]} \left(\sum_{j=3}^{N_{i,T}} \gamma_i(j-2)^{\frac{y}{1-y}}\right)^{1-y} && (4.22) \\
&\leq 2K + T^y \sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(N_{i,T}, \frac{y}{1-y}\right)^{1-y} \\
&\leq 2K + T^y K^y \left(\sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(N_{i,T}, \frac{y}{1-y}\right)\right)^{1-y} && (4.23) \\
&\leq 2K + T^y K \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil \frac{T}{K} \right\rceil, \frac{y}{1-y}\right)^{1-y}, && (4.24)
\end{aligned}
$$

where line (4.20) follows from the inequality $\min\{1, x\} \leq \min\{1, x\}^y \leq x^y$ for $y \in \left[0, \frac{1}{2}\right]$, line (4.21) follows from Hölder's inequality with powers $\frac{1}{y} \geq 1$ and $\frac{1}{1-y} \geq 1$ (since $y \in \left[0, \frac{1}{2}\right]$), line (4.22) is obtained from observing that $\sum_{j=3}^{N_{i,T}} (t_{i,j} - t_{i,j-1}) \leq T$ and $\gamma_i(t_{i,j-2}) \leq \gamma_i(j-2)$ from Assumption 2.2, line (4.23) follows from Jensen's inequality as $y \in \left[0, \frac{1}{2}\right]$ and observing:

$$
\begin{aligned}
\sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(N_{i,T}, \frac{y}{1-y}\right)^{1-y} &= K \sum_{i\in[K]} \frac{1}{K} \Upsilon_{\boldsymbol{\mu}}\left(N_{i,T}, \frac{y}{1-y}\right)^{1-y} \\
&\leq K^y \left(\sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(N_{i,T}, \frac{y}{1-y}\right)\right)^{1-y},
\end{aligned}
$$

and, finally, line (4.24) is obtained from Lemma A.3.

The final theorem statement is obtained by defining $q := \frac{y}{1-y} \in [0, 1]$ and substituting it

to the above equation. □

Stochastic Setting

We provide the regret analysis of `R-less-UCB` when we employ the exploration index:

$$B_i(t) \equiv \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) + \beta_i^{\texttt{R-less},h_{i,t}}(t),$$

where $\beta_i^{\texttt{R-less},h}(t)$ is the exploration bonus defined in Lemma 4.8 and $h_{i,t}$ is an arm and time-dependent window. The following result provides the regret bound, under particular choices of $h_{i,t}$ and δ_t.

Theorem 4.7. *Let $T \in \mathbb{N}$, then `R-less-UCB`(Algorithm 4.2) with $B_i(t) \equiv \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) + \beta_i^{\texttt{R-less},h_{i,t}}(t)$, $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0, 1/2)$, and $\delta_t = t^{-\alpha}$ for $\alpha > 2$, suffers an expected regret bounded, for every $q \in [0, 1]$, as:*

$$R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) \leq \mathcal{O}\Bigg(\frac{K}{\epsilon}(\sigma T)^{\frac{2}{3}} (\alpha \log T)^{\frac{1}{3}} + \frac{K T^{\frac{2q}{1+q}} (\log T)^{\frac{q}{1+q}}}{\epsilon(1-2\epsilon)} \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q \right)^{\frac{1}{1+q}} \Bigg).$$

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:

$$\mathcal{E}_{i,t} \coloneqq \left\{ \left| \widetilde{\mu}_i^{\texttt{R-less},h_{i,t}}(t) - \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) \right| \leq \beta_i^{\texttt{R-less},h_{i,t}}(t) \right\} \qquad \forall i \in [T],\, i \in [K].$$

We have to analyze the following expression:

$$R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i_t^*}(t) - \mu_{I_t}(t) \right],$$

where $i_t^* \in \operatorname{argmax}_{i \in [K]} \{\mu_i(t)\}$ for all $t \in [T]$.

We decompose according to the good events $\mathcal{E}_t$:

$$\begin{aligned} R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) &= \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right) \mathbb{1}\{\mathcal{E}_t\} \right] + \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right) \mathbb{1}\{\neg\mathcal{E}_t\} \right] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right) \mathbb{1}\{\mathcal{E}_t\} \right] + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\} \right], \end{aligned}$$

where we exploited $\mu_{i_t^*}(t) - \mu_{I_t}(t) \leq 1$ in the inequality.

Now, we bound the second summation, as done in Theorem 4.4:

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\neg\mathcal{E}_t\}\right] \leq 1 + \frac{2K}{\alpha - 2}.$$

From now on, we will proceed the analysis under the good event $\mathcal{E}_t$, recalling that $B_i(t) \equiv \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) + \beta_i^{\texttt{R-less},h_{i,t}}(t)$.

Let $t \in [T]$, and we exploit the optimism, i.e., $B_{i_t^*}(t) \leq B_{I_t}(t)$:

$$\begin{aligned}\mu_{i^*}(t) - \mu_{I_t}(t) + B_{I_t}(t) - B_{I_t}(t) &\leq \min\left\{1, \underbrace{\mu_{i_t^*}(t) - B_{i_t^*}(t)}_{\leq 0} + B_{I_t}(t) - \mu_{I_t}(t)\right\} \\ &\leq \min\left\{1, B_{I_t}(t) - \mu_{I_t}(t)\right\}.\end{aligned}$$

Now, we work on the term inside the minimum:

$$\begin{aligned}B_{I_t}(t) - \mu_{I_t}(t) &= \widehat{\mu}_{I_t}^{\texttt{R-less},h_{I_t,t}}(t) + \beta_{I_t}^{\texttt{R-less},h_{I_t,t}}(t) - \mu_{I_t}(t) && (4.25)\\ &\leq \underbrace{\widetilde{\mu}_{I_t}^{\texttt{R-less},h_{I_t,t}}(t) - \mu_{I_t}(t)}_{\text{(a)}} + \underbrace{2\beta_{I_t}^{\texttt{R-less},h_{I_t,t}}(t)}_{\text{(b)}}, && (4.26)\end{aligned}$$

where line (4.25) follows from the definition of $B_i(t)$ and line (4.26) from the good event $\mathcal{E}_t$.

We proceed decomposing over the arms, starting with (a):

$$\sum_{t=1}^{T} \min\left\{1, \widetilde{\mu}_{I_t}^{\texttt{R-less},h_{I_t,t}}(t) - \mu_{I_t}(t)\right\} \tag{4.27}$$

$$\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\left\{1, \widetilde{\mu}_{i}^{\texttt{R-less},h_{i,t_{i,j}}}(t_{i,j}) - \mu_i(t_{i,j})\right\}$$

$$\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\Bigg\{1, \frac{(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)(t_{i,j-1} - t_{i,j-2h_{i,t}+1})}{2h_{i,t}}\gamma_i(t_{i,(j-1)-2h_{i,t_{i,j}}+1})\Bigg\} \tag{4.28}$$

$$\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{T^2}{\lfloor\epsilon(j-1)\rfloor}\gamma_i(t_{i,j-2\lfloor\epsilon(j-1)\rfloor})\right\} \tag{4.29}$$

$$\leq 2K + \sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \min\left\{1, \frac{T^2}{\lfloor\epsilon(j-1)\rfloor}\gamma_i(\lfloor(1-2\epsilon)j\rfloor)\right\} \tag{4.30}$$

$$\leq 2K + T^{2z}\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}} \left(\frac{\gamma_i(\lfloor(1-2\epsilon)j\rfloor)}{\lfloor\epsilon(j-1)\rfloor}\right)^z \tag{4.31}$$

$$\leq 2K + T^{2z}\sum_{i\in[K]} \left(\sum_{j=3}^{N_{i,T}} \frac{1}{\lfloor\epsilon(j-1)\rfloor}\right)^z \left(\sum_{j=3}^{N_{i,T}} \gamma_i(\lfloor(1-2\epsilon)j\rfloor)^{\frac{z}{1-z}}\right)^{1-z} \tag{4.32}$$

$$\leq 2K + T^{2z}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{i\in[K]} \left(\sum_{j=\lfloor 2\epsilon\rfloor}^{\lfloor\epsilon(N_{i,T}-1)\rfloor} \frac{1}{j}\right)^z \left(\sum_{j=\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor} \gamma_i(j)^{\frac{z}{1-z}}\right)^{1-z} \tag{4.33}$$

$$\leq 2K + T^{2z}(1+\log(\epsilon T))^z\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{i\in[K]} \left(\sum_{j=\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor} \gamma_i(j)^{\frac{z}{1-z}}\right)^{1-z} \tag{4.34}$$

$$\leq 2K + T^{2z}(1+\log(\epsilon T))^z\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(\lfloor(1-2\epsilon)N_{i,T}\rfloor, \frac{z}{1-z}\right)^{1-z}$$

$$\leq 2K + T^{2z}(1+\log(\epsilon T))^z\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil K^z\left(\sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(\lfloor(1-2\epsilon)N_{i,T}\rfloor, \frac{z}{1-z}\right)\right)^{1-z} \tag{4.35}$$

$$\leq 2K + T^{2z}(1+\log(\epsilon T))^z\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil K\Upsilon_{\boldsymbol{\mu}}\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil, \frac{z}{1-z}\right)^{1-z}, \tag{4.36}$$

where line (4.28) follows from the bias bound of Lemma 4.7, line (4.29) is obtained from bounding $(2t_{i,j} - 2(j-1) + h_{i,t_{i,j}} - 1)(t_{i,j-1} - t_{i,j-2h_{i,t}+1}) \leq 2T^2$ and using the definition of $h_{i,t}$, line (4.30) derives from observing that $\gamma_i(t_{i,j}) \leq \gamma_i(j)$ for Assumption 2.2 and having

bounded the floor analogously as done in Theorem 4.4, line (4.31) from the inequality $\min\{1,x\} \leq \min\{1,x\}^z \leq x^z$ for $z \in [0,1/2]$, line (4.32) is obtained from Hölder's inequality with exponents $\frac{1}{z} \geq 1$ and $\frac{1}{1-z} \geq 1$ respectively, line (4.33) is an application of Lemma A.2 to independently to both inner summations, line (4.34) derives from bounding the harmonic sum, i.e.,$\sum_{\lfloor 2\epsilon \rfloor}^{\lfloor \epsilon(N_{i,T}-1) \rfloor} \frac{1}{j} \leq 1 + \log(\epsilon(N_{i,T}-1)) \leq 1 + \log(\epsilon T)$, line (4.35) follows from Jensen's inequality, line (4.36) is obtained from Lemma A.3.

By recalling $q = \frac{z}{1-z} \in [0,1]$, we obtain:

$$2K + T^{\frac{2q}{1+q}}(1+\log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil K \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q \right)^{\frac{1}{1+q}}.$$

Concerning the term (b), we recall that $\beta_{I_t}^{\texttt{R-less},h_{I_t,t}}(t)$ equals the bonus term used in the rested setting and, consequently from Theorem 4.4:

$$\sum_{t=1}^{T} \min\left\{1, 2\beta_{I_t}^{\texttt{R-ed},h_{I_t,t}}(t,\delta_t)\right\} \leq K\left(3+\frac{1}{\epsilon}\right) + \frac{3K}{\epsilon}(2\sigma T)^{\frac{2}{3}}(10\alpha \log T)^{\frac{1}{3}}.$$

Putting all together, we obtain:

$$\begin{aligned}
&R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) \\
&\quad \leq 1 + \frac{2K}{\alpha-2} + 5K + \frac{K}{\epsilon} + \frac{3K}{\epsilon}(2\sigma T)^{\frac{2}{3}}(10\alpha\log T)^{\frac{1}{3}} \\
&\qquad + T^{\frac{2q}{1+q}}(1+\log(\epsilon T))^{\frac{q}{1+q}} \left\lceil \frac{1}{\epsilon} \right\rceil \left\lceil \frac{1}{1-2\epsilon} \right\rceil K \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q \right)^{\frac{1}{1+q}} \\
&\qquad = \mathcal{O}\left(\frac{K}{\epsilon}(\sigma T)^{\frac{2}{3}}(\alpha \log T)^{\frac{1}{3}} + \frac{K T^{\frac{2q}{1+q}}(\log T)^{\frac{q}{1+q}}}{\epsilon(1-2\epsilon)} \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil (1-2\epsilon)\frac{T}{K} \right\rceil, q \right)^{\frac{1}{1+q}} \right).
\end{aligned}$$

□

5 | Numerical Simulations

In this chapter, we provide the results of several experiments, performed both on synthetically generated and real-world data. The main goal is to describe and analyze the different performances obtained by the state-of-the-art algorithms (introduced in Chapter 3) w.r.t. the novels `R-less-UCB` and `R-ed-UCB`. Due to the absence of suitable benchmarks for the rested rising bandit problem, we decided to use the state-of-the-art restless algorithms as a term of comparison in the rested setting too.

The chapter is organized as follows. First of all, we explain the rational behind the numerical experiments, then we procede explaining the results obtained by `R-less-UCB` in the restless setting (Section 5.2), and by `R-ed-UCB` in the rested setting (Section 5.3). Finally we present an experiment performed on a real-world online model selection task (Section 5.3.1), showing it can be solved using `R-ed-UCB`.

5.1. Metodology

Rising Rewards We evaluated the algorithms' performances over different bandits whose arms' payoff functions follow randomly chosen synthetical functions. Particularly we created the payoff function of each arm $\mu_i(x)$ by randomly sampling a function f_i from one of the following families:

$$F_{\text{exp}} = \left\{ f(x) = c\left(1 - e^{-ax}\right) \right\},$$
$$F_{\text{poly}} = \left\{ f(x) = c\left(1 - b(x + b^{1/\rho})^{-\rho}\right) \right\},$$

where $a, c, \rho \in (0, 1]$ and $b \in \mathbb{R}_{\geq 0}$ are parameters, whose values have been selected randomly. By construction all functions $f \in F_{\text{exp}} \cup F_{\text{poly}}$ satisfy Assumptions 2.1 and 2.2.

The two families allow having models of functions which behave differently, indeed the ones coming from the first family, F_{exp} (exponential functions), display a fast increase, while the ones from F_{poly} (polynomial functions) have a slower growth rate $\gamma_i(x)$; hence, different cumulative increments Υ_{μ} are modeled. Having fixed the evolution of each arm payoff as $\mu_i(t) = f_i(t)$ or $\mu_i(N_{i,t}) = f_i(N_{i,t})$ in the restless and in the rested setting

respectively, the stochasticity is implemented by adding a Gaussian noise with $\sigma = 0.1$. To summarize, the reward observed by the agent is obtained, at each round, by sampling a Gaussian distribution $R_t \sim \mathcal{N}(f_i(x), \sigma)$ being $x \coloneqq t$ in the restless setting and $x \coloneqq N_{i,t}$ in the rested.

Rising Algorithms For both `R-less-UCB` and `R-ed-UCB`, the window is set as $h_{i,t} = \lfloor N_{i,t-1}/4 \rfloor$ (according to Theorems 4.4 and 4.7).

Benchmarks Among the algorithms for non-stationary MABs (Chapter 3), we considered as baselines: `Rexp3` (Algorithm 3.4), to understand how a variation budget approach works in a rising setting, `KL-UCB` (Algorithm 1.2), to analyze the differences with a stationary scenario, `Ser4` (Algorithm 3.5), to highlight a best-arm-switch approach, and sliding-window algorithms such as `SW-UCB` (Algorithm 3.1), `SW-KL-UCB` (Algorithm 3.2), and `SW-TS` (Algorithm 3.3), to see how algorithms that are generally able to deal with non-stationary restless settings perform in a rising scenario. The parameters for all the baseline algorithms have been set as recommended by the corresponding authors (further details are provided in Appendix C).

Evaluation The term of comparison is the empirical cumulative regret $\widehat{R}_{\boldsymbol{\mu}}(\pi, t)$, i.e., the empirical counterpart of the expected cumulative regret $R_{\boldsymbol{\mu}}(\pi, t)$ at round t, averaged over 100 independent runs.

5.2. Restless

To evaluate `R-less-UCB` in the restless setting, we run the previously introduced algorithms on a problem with $K = 15$ arms over a time horizons of $T = 200,000$ rounds. The randomly generated functions for this experiment are shown in Figure 5.1. Clearly in the restless setting the evolution of the functions depend only on the round t.

Figure 5.2 shows the empirical cumulative regret $\widehat{R}_{\boldsymbol{\mu}}(\pi, t)$ obtained by averaging 100 independent runs of each algorithm, with the related 95% confidence intervals. Some considerations: the results show that `SW-TS` is the algorithm that achieves the lowest regret at the horizon, even though its performance at the beginning is worse than the other algorithms, included `R-less-UCB`. As commonly happening in practice, Thompson Sampling based approaches tend to outperform Upper Confidence Bound ones. `R-less-UCB` displays the second-best curve overall and achieves the best performance among the UCB-like algorithms.

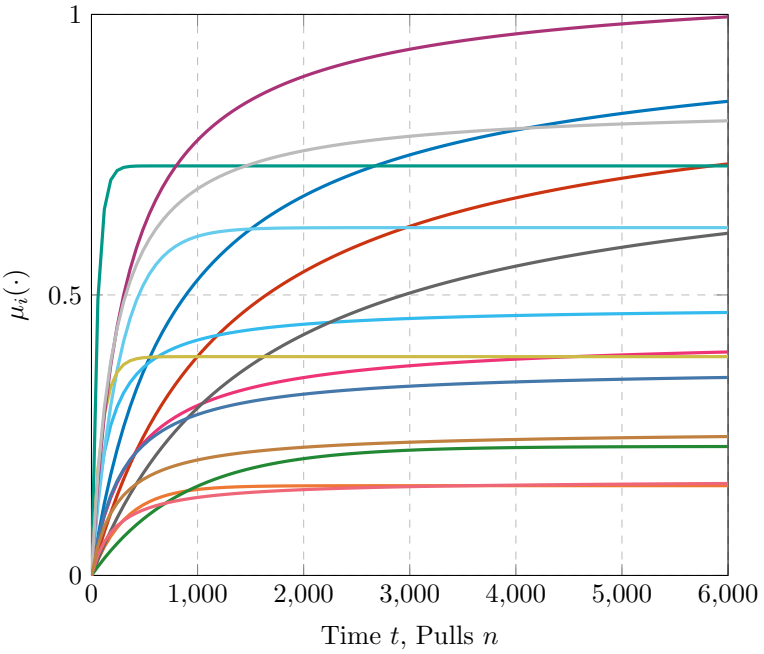


Figure 5.1: 15 arms bandit setting: first 6000 rounds/pulls of the payoff functions.

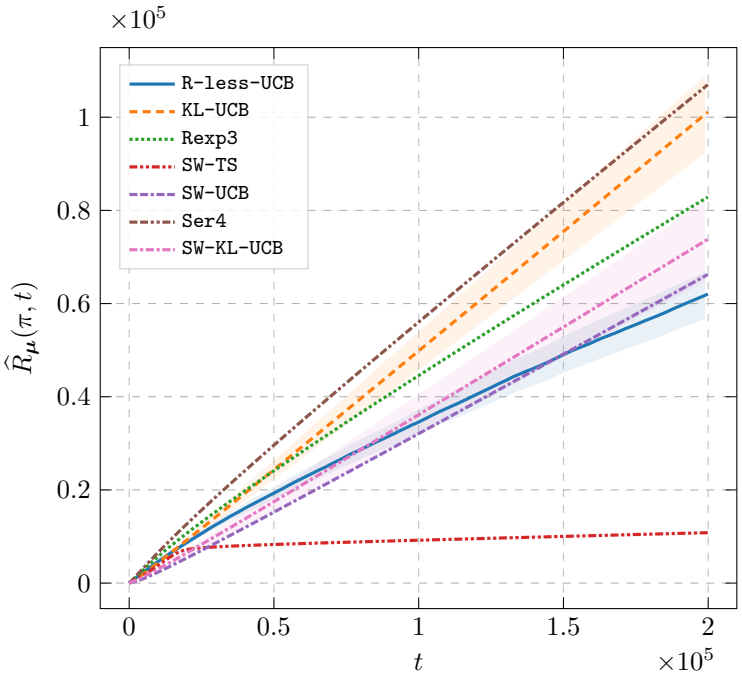


Figure 5.2: 15 arms bandit setting: cumulative regret in the **restless** scenario with horizon $T = 200,000$ (100 runs 95% c.i.).

Ranking

In order not to limit our testing to a specific instance of a bandit, we have generated a wider range of bandit environments. The performance of the algorithms was evaluated with the early-discussed approach over 50 different bandits with $K \in \{2, \dots, 15\}$ randomly generated arms over a time horizon of $T = 200,000$. The empirical regret obtained by each algorithm in each scenario is the result of the average of 10 independent experiments. We assigned a score to each algorithm w.r.t. the other competitors in that scenario, i.e., 1 for the best performing algorithm (lowest regret) and 7 for the one with largest regret, in order to create a ranking based on the results obtained in the 50 tests. The results are summarized in Table 5.1. `R-less-UCB` achieves a worse-than-average performance, probably influenced by the characteristics of the randomly generated bandits. Due to this unsatisfactory results, we propose a slight modification of `R-less-UCB`, recalling the upper bound which intuitively followed from the derivation in Section 4.2.1:

$$\widehat{\mu}_i^{\texttt{R-less-H},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(R_{t_{i,l}} + (t - t_{i,l}) \frac{R_{t_{i,l}} - R_{t_{i,l-h}}}{t_{i,l} - t_{i,l-h}} \right).$$

We call this new algorithm `R-less-UCB-H` to denote it is an heuristic method; indeed, while the performance of the heuristic seems good in practice (it achieves the best overall result), its downside is that the theoretical guarantees on the regret will have to be reconsidered.

5.3. Rested

We repeated the specific 15-arms bandit experiment designed for the restless case to evaluate `R-ed-UCB` in the rested setting; Figure 5.1 presents each reward function $\mu_i(n)$, where the evolution is controlled not by the round t but by the pulls $n \equiv N_{i,t}$.

In Figure 5.3, we have plotted the empirical cumulative regret $\widehat{R}_{\boldsymbol{\mu}}(\pi, t)$ obtained by averaging 100 independent runs of each algorithm, with the related 95% confidence intervals, when the horizon of the problem is set to $T = 200,000$. `SW-TS` is confirmed as the best algorithm at the end of the time horizon in the rested setting too, although other algorithms like `SW-UCB` and `SW-KL-UCB` suffer less regret at the beginning of the learning process. `R-ed-UCB` presents a behaviour similar to `SW-KL-UCB`, paying the price of the initial exploration, but it manages to achieve the second best performance overall. `R-ed-UCB` has a longer exploration phase w.r.t. the others.

Moreover, it is important to notice that, all the algorithms besides `R-ed-UCB` are designed for a restless setting and, consequently, are not endowed with any theoretical guarantee

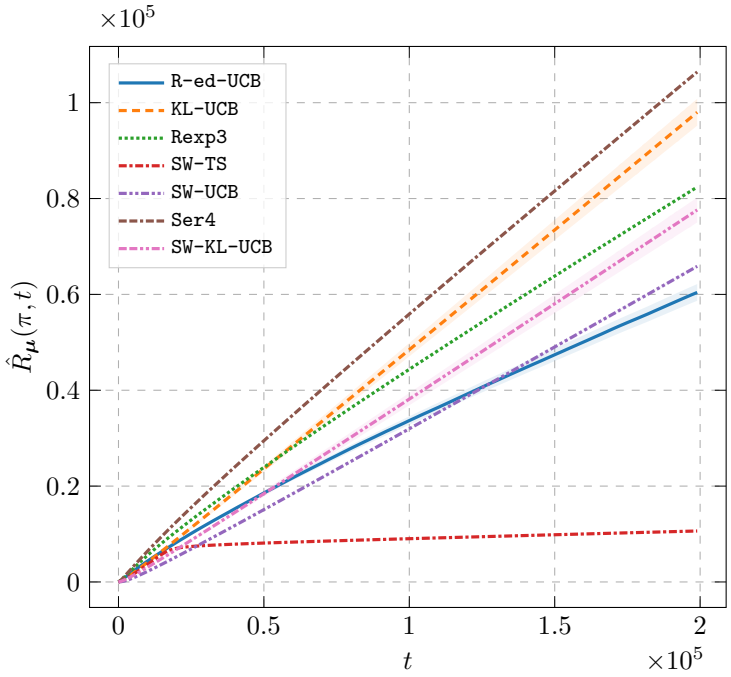


Figure 5.3: 15 arms bandit setting: cumulative regret in the **rested** scenario with horizon $T = 200,000$ (100 runs 95% c.i.).

on the regret in the rested scenario. In order to clarify this fact, we designed a specific example of a 2-arms rising rested bandit in which the optimal arm is hidden until it is pulled a sufficient number of times (linear in T). This particular scenario will highlight the different reactions of the algorithms: the payoff functions, fulfilling Assumptions 2.1 and 2.2 by construction, are shown in Figure 5.4a and the algorithms' empirical regrets in Figure 5.4b, when the horizon is set to $T = 200,000$. Notice that in this particular experiment the expected (instantaneous) regret may be negative due to the fact that the *oracle constant* policy $\pi^{c}_{\boldsymbol{\mu},T}$ keeps pulling the *overall* best arm, which is suboptimal in the first $\approx 20,000$ rounds, hence algorithms which pull the other arm in this initial rounds achieve an instantaneous performance better than the oracle. In the initial $\approx 20,000$ rounds `R-ed-UCB` behaves similarly to the baselines, but in the second part of the learning process it clearly outperforms all the other options. Indeed the other algorithms did not notice the switch in the best arm which occurs when the arms are pulled at least $19T/100$ times each, since they are not prompt to detect such a change. It is possible to notice that the regret slopes of `Rexp3` and `Ser4` are the first to decrease at $t \approx 40,000$, meaning that they are somehow reacting to the best arm switch; their behaviour is later followed by `SW-TS` at round $t \approx 100,000$, while the remaining baselines suffer a linear regret: `KL-UCB` does not have forgetting mechanisms, while the sliding window of `SW-KL-UCB` and `SW-UCB` should be tuned accordingly to the features of the problem.

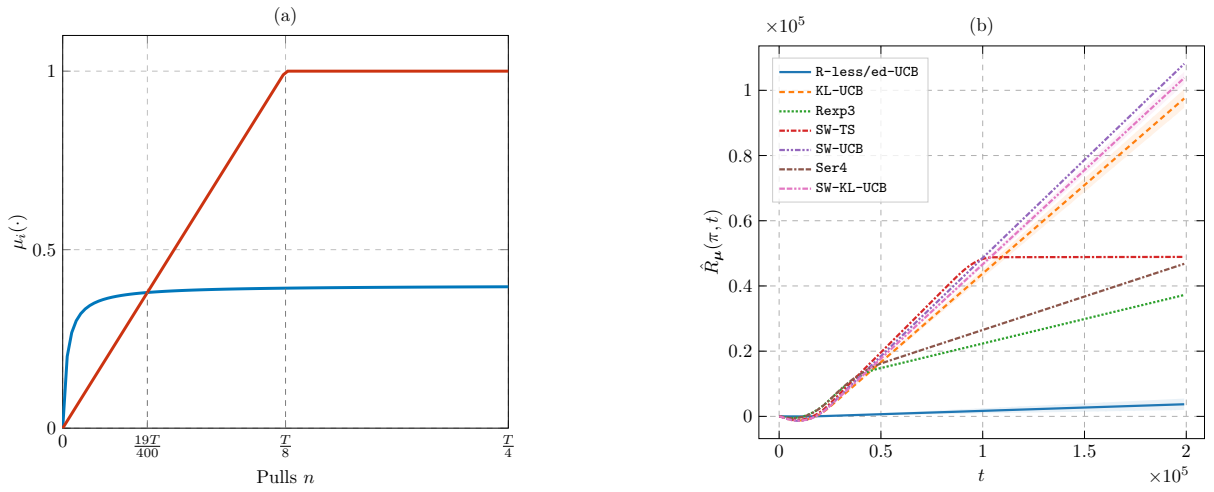


Figure 5.4: 2 arms **rested** bandit setting: (a) payoff functions, (b) cumulative regret when $T = 200,000$ (100 runs, 95% c.i.).

Table 5.1: **Ranking** of the algorithms (50 bandits, 10 runs, 95% c.i. in brackets).

Algorithm	Restless	Restless Heuristic	Rested
`R-ed-UCB`	–	–	4.98 (0.34)
`R-less-UCB`	5.14 (0.38)	–	–
`R-less-UCB-H`	–	1.90 (0.30)	–
`KL-UCB`	2.54 (0.34)	2.46 (0.31)	2.56 (0.43)
`Rexp3`	5.20 (0.26)	6.08 (0.16)	5.10 (0.26)
`SW-TS`	2.86 (0.39)	4.76 (0.19)	2.84 (0.35)
`SW-UCB`	2.58 (0.47)	3.08 (0.30)	2.12 (0.44)
`Ser4`	6.60 (0.28)	6.66 (0.18)	6.84 (0.15)
`SW-KL-UCB`	3.08 (0.45)	3.06 (0.48)	3.56 (0.38)

Ranking

We applied the same methodology used in the restless ranking (Section 5.2) in order to evaluate the performances of all the algorithms over many different rested rising bandit problems and draw up a leaderboard. The results are summarized in Table 5.1. In the rested case, `R-ed-UCB` is among the worst algorithms, placing 4.18 on average. This is due to the fact that on average the algorithm is not superior to the baselines, but it comes with theoretical guarantees in the rising rested settings, while the other algorithms do not, as already shown in the 2-arms experiment.

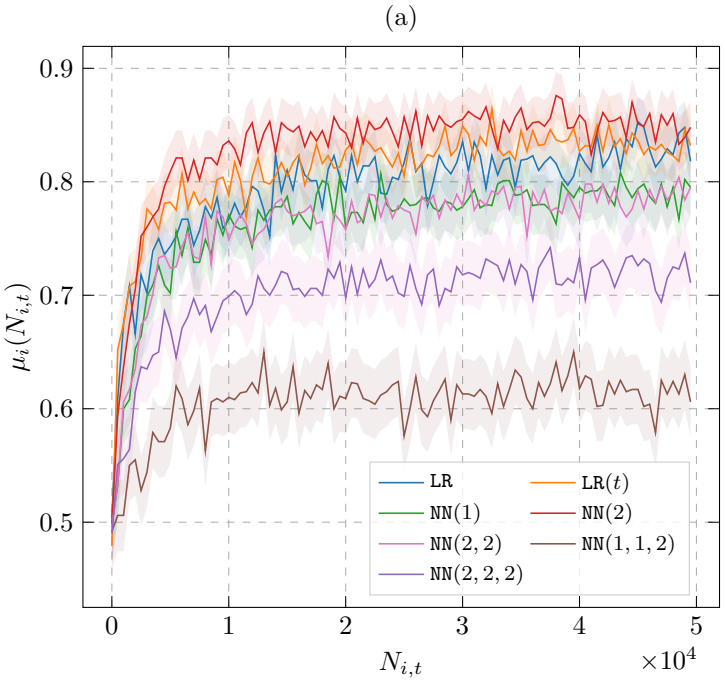


Figure 5.5: IMDB experiment: empirical learning curves of the classification algorithms (arms).

5.3.1. IMDB dataset

We tested the performance of `R-ed-UCB` on a real-world online model selection problem. We extracted 50,000 reviews of movies from the IMDB dataset and preprocessed the data, as previously done by Maas et al. [37], separating the reviews with a score higher than 5 (positive class) with the others (negative class), in order to obtain a binary classification problem; each review $x_t \in \mathbb{R}^d$, being $d = 10,000$ the number of features, consists of the frequencies of the most common english words.

For the classification task we decided to employ:

- 2 Online Logistic Regression (`OLR`) methods with different schemes used for the learning rate λ_t;
- 5 Neural Networks (`NNs`) different in terms of shape and number of neurons

We will refer to the above algorithms as "base algorithms". For the `OLR` algorithms, we adopt a decreasing scheme for the first one, denoted with `LR`(t), $\lambda_t = \frac{\beta}{t}$, and a constant learning rate for the latter `LR`, $\lambda_t = \beta$. For the `NNs`, their activation functions are the rectified linear unit, i.e., $relu(x) = \max(0, x)$, their learning rate is constant $\alpha = 0.001$ and their optimization method for fitting is the "Adam" stochastic gradient optimizer.

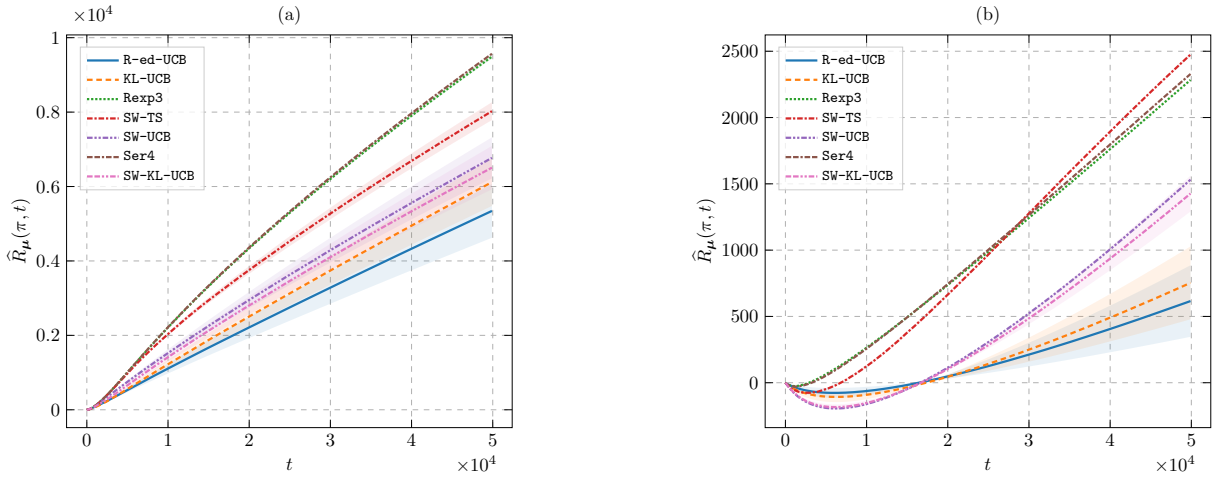


Figure 5.6: IMDB experiment: cumulative regret in the setting with $K = 7$ arms (a) and with $K = 2$ arms (b) (30 runs 95% c.i.).

Two of the chosen nets have only one hidden layer, with 1 and 2 neurons, respectively, the third net has 2 hidden layer, with 2 neurons each, and two nets have 3 layers with $2, 2, 2$ and $1, 1, 2$ neurons, respectively. We refer to a specific `NN` denoting in curve brackets the cardinalities of the layers, e.g., the one having 2 layer with 2 neurons each is denoted by `NN`$(2, 2)$.

We created a bandit environment in which each base algorithm is an arm of the bandit. At each round t:

- the agent decides to pull arm I_t, i.e., decides to perform the classification task with a specific base algorithm;
- a sample x_t of the IMDB dataset is randomly selected and supplied to the base algorithm corresponding to arm I_t;
- the base algorithm classifies the sample, i.e., provides the prediction $\hat{y}_t \in \{0, 1\}$ for the selected sample x_t;
- the environment generates the reward comparing the prediction $\hat{y}_t$ to the target class y_t using the following function $R_t = 1 - |y_t - \hat{y}_t|$, i.e., 1 for a correct prediction and 0 for a wrong one;
- the base algorithm linked to arm I_t is updated using (x_t, y_t).

Being the base algorithms trained only if the arm they are associated with is selected, this is a problem which belongs to the rested bandit scenario. Moreover we expect the performance of each base algorithm to improve the more the it is trained, hence the rewards,

on average, increase. To this purpose, we generated the average learning curves (i.e., the value of the payoff $\mu_i(n)$) of the base algorithms on the IMDB dataset by averaging $1,000$ independent runs in which each classification algorithm is sequentially fed with all the available $50,000$ samples. As it is possible to notice from Figure 5.5, the average learning curves are, qualitatively speaking, increasing and concave, however Assumptions 2.1 and 2.2 are not globally satisfied.

We evaluated the performance of the MAB algorithms averaging 30 independent runs both when all the 7 classification algorithms were available and when only the 2 logistic regression algorithms were present, obtaining similar results. Particularly, the empirical regret obtained by each bandit algorithm in those scenarios is plotted in Figure 5.6.

It is possible to notice how `R-ed-UCB` outperforms the considered baselines, confirming the results obtained in the synthetical simulations. We also notice that good performances are obtained by `KL-UCB`, while other algorithms degenerate to linear regret.

6 | Conclusions & Future Developments

Main Results This work studied the stochastic Multi Armed Bandit problem when the payoff are non-decreasing concave functions that evolve either for time passing (restless setting) or when pulling the corresponding arm (rested setting). The results obtained are the following:

- we have formally introduced the stochastic rising bandit framework in the restless and rested formulation;
- we have shown that the rested bandit problem with non-decreasing payoffs is *non-learnable* unless additional assumptions on the payoff functions are enforced, such as the concavity;
- we have presented two optimistic algorithms, `R-less-UCB` and `R-ed-UCB`, for the rising restless and rested bandits respectively; moreover the former is the first algorithm designed for stochastic rested rising bandits;
- we have shown that both algorithms suffer an expected regret composed of an instance-dependent term $\Upsilon_{\boldsymbol{\mu}}(T, q)$ and an instance-independent component, which in the worst case can be bounded by $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$;
- we have illustrated, using both synthetic and real-world data, the advantages of the proposed approaches w.r.t. the state-of-the-art algorithms for the non-stationary (restless) bandits.

Further Applications Driven by online model selection tasks, a natural improvement of our work is in the best-arm identification setting, i.e., a scenario in which the goal is not to minimize the cumulative regret, but to spot, in the fastest way possible and with high probability, which arm is the best. Another reasonable application, which surely deserves future studies, is an online model selection setting in which the base algorithms cooperates together to find a solution, e.g., through the means of a shared vector of parameters.

Future Research A straightforward continuation of our work consists in the theoretical study of the stochastic rising bandit learning problem in order to derive suitable regret *lower bounds* for both the restless and the rested setting. Such bounds would be crucial to understand the possibility to improve `R-less-UCB` and `R-ed-UCB` regret guarantees, as much as for the future development of rising bandit algorithms. Other studies should fill the gap with the lack of rested bandit algorithms, both developing solutions for the rising scenario and compare the results and guarantees with `R-ed-UCB`.

Bibliography

[1] Y. Abbasi-Yadkori, A. Pacchiano, and M. Phan. Regret balancing for bandit and RL model selection. *CoRR*, abs/2006.05491, 2020. URL `https://arxiv.org/abs/2006.05491`.

[2] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In S. Kale and O. Shamir, editors, *Proceedings of the Conference on Learning Theory (COLT)*, volume 65, pages 12–38, 2017.

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

[4] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the international conference on Artificial Intelligence and Statistics (AISTATS)*, pages 99–107. PMLR, 2013.

[5] R. Allesiardo, R. Féraud, and O.-A. Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4): 267–283, 2017.

[6] R. Arora, T. V. Marinov, and M. Mohri. Corralling stochastic bandit algorithms. In A. Banerjee and K. Fukumizu, editors, *Proceedings of the international conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 2116–2124, 2021.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the IEEE annual symposium on Foundations of Computer Science (FOCS)*, pages 322–331, 1995.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[9] P. Auer, P. Gajane, and R. Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In A. Beygelzimer and D. Hsu, editors,

Proceedings of the Conference on Learning Theory (COLT), volume 99, pages 138–158, 2019.

[10] M. Aziz, E. Kaufmann, and M.-K. Riviere. On multi-armed bandit designs for dose-finding clinical trials. *Journal of Machine Learning Research*, 22(14):1–38, 2021.

[11] O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Proceedings of the conference on Neural Information Processing Systems (NeurIPS)*, volume 27, pages 199–207, 2014.

[12] L. Besson, E. Kaufmann, O.-A. Maillard, and J. Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. 2019. URL `https://arxiv.org/abs/1902.01575`.

[13] G. Bresler, G. H. Chen, and D. Shah. A latent source model for online collaborative filtering. *Advances in neural information processing systems*, 27, 2014.

[14] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems 21 (NeurIPS)*, pages 201–208, 2008.

[15] Y. Cao, Z. Wen, B. Kveton, and Y. Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The international conference on Artificial Intelligence and Statistics (AISTATS)*, pages 418–427, 2019.

[16] L. Cella, M. Pontil, and C. Gentile. Best model identification: A rested bandit formulation. In M. Meila and T. Zhang, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 1362–1372, 2021.

[17] Y. Chen, C. Lee, H. Luo, and C. Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 99, pages 696–726, 2019.

[18] S.-C. Chow and M. Chang. Adaptive design methods in clinical trials–a review. *Orphanet journal of rare diseases*, 3(1):1–13, 2008.

[19] R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 521–529, 2014.

[20] J. L. Doob and J. L. Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.

[21] A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits

and beyond. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 359–376, 2011.

[22] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. 2008. URL `https://arxiv.org/abs/0805.3415`.

[23] A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the international conference on Algorithmic Learning Theory (ALT)*, pages 174–188, 2011.

[24] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

[25] H. Heidari, M. J. Kearns, and A. Roth. Tight policy regret bounds for improving and decaying bandits. In S. Kambhampati, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1562–1570, 2016.

[26] M. N. Katehakis and H. Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92 (19):8584, 1995.

[27] R. Kleinberg and N. Immorlica. Recharging bandits. In *Proceedings of the IEEE annual symposium on Foundations of Computer Science (FOCS)*, pages 309–319, 2018.

[28] T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

[29] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[30] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[31] J. Le Ny, M. Dahleh, and E. Feron. Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In *2008 American Control Conference*, pages 4220–4225. IEEE, 2008.

[32] N. Levine, K. Crammer, and S. Mannor. Rotting bandits. In *Proceedings of the conference on Neural Information Processing Systems (NeurIPS)*, pages 3074–3083, 2017.

[33] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to

personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[34] S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016.

[35] Y. Li, J. Jiang, J. Gao, Y. Shao, C. Zhang, and B. Cui. Efficient automatic CASH via rising bandits. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 4763–4771, 2020.

[36] F. Liu, J. Lee, and N. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

[37] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the annual meeting of the association for computational linguistics: Human Language Technologies (HLT)*, pages 142–150, 2011.

[38] Y. Mintz, A. Aswani, P. Kaminsky, E. Flowers, and Y. Fukuoka. Nonstationary bandits with habituation and recovery dynamics. *Operations Research*, 68(5):1493–1516, 2020.

[39] A. Nuara, F. Trovo, N. Gatti, and M. Restelli. A combinatorial-bandit algorithm for the online joint bid/budget optimization of pay-per-click advertising campaigns. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[40] A. Nuara, F. Trovò, N. Gatti, and M. Restelli. Online joint bid/daily budget optimization of internet advertising campaigns. *Artificial Intelligence*, 305:103663, 2022.

[41] R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless markov bandits. In *Proceedings of the international conference on Algorithmic Learning Theory (ALT)*, pages 214–228, 2012.

[42] A. Pacchiano, C. Dann, C. Gentile, and P. L. Bartlett. Regret bound balancing and elimination for model selection in bandits and RL. *CoRR*, abs/2012.13045, 2020. URL `https://arxiv.org/abs/2012.13045`.

[43] A. Pacchiano, M. Phan, Y. Abbasi-Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvári. Model selection in contextual stochastic bandit problems. In *Proceedings of the conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[44] M. Parvin and M. R. Meybodi. Mabrp: A multi-armed bandit problem-based energy-aware routing protocol for wireless sensor network. In *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*, pages 464–468. IEEE, 2012.

[45] C. Pike-Burke and S. Grunewalder. Recovering bandits. *Proceedings of the conference on Neural Information Processing Systems (NeurIPS)*, 32:14122–14131, 2019.

[46] G. Re, F. Chiusano, F. Trovò, D. Carrera, G. Boracchi, and M. Restelli. Exploiting history data for nonstationary multi-armed bandit. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 51–66. Springer, 2021.

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

[48] Y. Russac, C. Vernade, and O. Cappé. Weighted linear bandits for non-stationary environments. In *Proceedings of the conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[49] D. Sauré and A. Zeevi. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

[50] E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

[51] J. Seznec, A. Locatelli, A. Carpentier, A. Lazaric, and M. Valko. Rotting bandits are no harder than stochastic ones. In *Proceedings of the international conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 2564–2572, 2019.

[52] J. Seznec, P. Ménard, A. Lazaric, and M. 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, pages 3784–3794, 2020.

[53] C. Tekin and M. Liu. Online learning of rested and restless bandits. *IEEE Transaction on Information Theory*, 58(8):5588–5611, 2012.

[54] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[55] F. Trovò, S. Paladino, M. Restelli, and N. Gatti. Improving multi-armed bandit al-

gorithms in online pricing settings. *International Journal of Approximate Reasoning*, 98:196–235, 2018.

[56] F. Trovò, S. Paladino, M. Restelli, and N. Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

A | Proofs and Derivations

Here we provide the proofs of the results presented in Chapter 4. First of all we present detailed derivations of the upper bounds introduced in the main work (Sections 4.1.1, 4.2.1), then we procede introducing Thecnical Lemmas used to derive the regret bounds of Sections 4.1.3, 4.2.3.

A.1. Proofs Rested Setting (Section 4.1)

Theorem 4.1 (Constant Policy Optimality [25]). *Let $\pi^c_{\boldsymbol{\mu},T}$ be the* oracle constant *policy:*

$$\pi^c_{\boldsymbol{\mu},T}(t) \in \underset{i\in[K]}{\operatorname{argmax}} \left\{ \sum_{l\in[T]} \mu_i(N_{i,l-1}) \right\}, \quad \forall t \in [T].$$

Then, $\pi^c_{\boldsymbol{\mu},T}$ is optimal for the rested non-decreasing bandits (i.e., under Assumption 2.1).

Proof. The proof is reported in Proposition 1 of Heidari et al. [25]. □

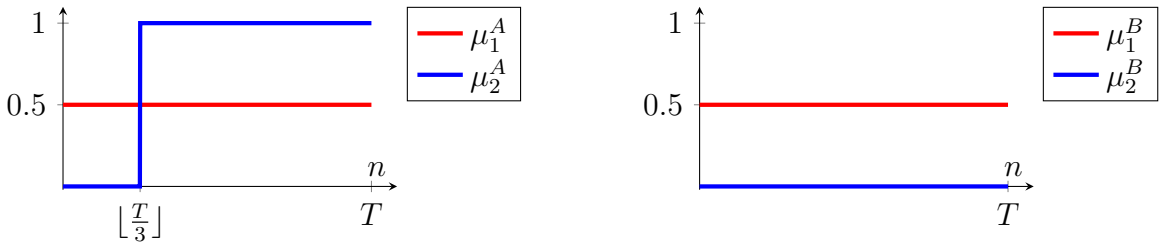
Lemma A.1 (Non Learnable). *In the noiseless ($\sigma = 0$) setting, there exists a 2-armed non-decreasing non-concave rested bandit such that any learning policy π suffers regret:*

$$R_{\boldsymbol{\mu}}(\pi, T) \geq \left\lfloor \frac{T}{12} \right\rfloor.$$

Proof. Let $\boldsymbol{\mu}^A$ and $\boldsymbol{\mu}^B$ be two non-concave non-decreasing rested bandits, defined as:

$$\begin{aligned}
&\mu^A_1(n) = \mu^B_1(n) = \frac{1}{2}, \\
&\mu^A_2(n) = \begin{cases} 0 & \text{if } n \leq \lfloor \frac{T}{3} \rfloor \\ 1 & \text{otherwise} \end{cases}, \\
&\mu^B_2(n) = 0.
\end{aligned}$$

It is clear that for $\boldsymbol{\mu}^A$ the optimal arm is 2, whereas for bandit $\boldsymbol{\mu}^B$ the optimal arm is 1, having optimal performance respectively $J^*_{\boldsymbol{\mu}^A}(T) = \lceil \frac{2}{3}T \rceil$ and $J^*_{\boldsymbol{\mu}^B}(T) = \frac{T}{2}$.



Let π be an arbitrary policy. Since the learner will receive the same rewards for both bandits until at least $\lfloor \frac{T}{3} \rfloor$, we have:

$$\pi\left(\mathcal{H}_{\lfloor \frac{T}{3} \rfloor}(\boldsymbol{\mu}^A)\right) = \pi\left(\mathcal{H}_{\lfloor \frac{T}{3} \rfloor}(\boldsymbol{\mu}^B)\right) \implies \mathbb{E}_{\boldsymbol{\mu}^A}\left[N_{1,\lfloor \frac{T}{3} \rfloor}\right] = \mathbb{E}_{\boldsymbol{\mu}^B}\left[N_{1,\lfloor \frac{T}{3} \rfloor}\right] =: n_1.$$

Let us now compute the performance of policy π in the two bandits and the corresponding regrets. Let us start with $\boldsymbol{\mu}^A$:

$$J_{\boldsymbol{\mu}^A}(\pi, T) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] + \max\left\{0, \mathbb{E}_{\boldsymbol{\mu}^A}[N_{2,T}] - \left\lfloor \frac{T}{3} \right\rfloor\right\} \tag{A.1}$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] + \max\left\{0, \left\lceil \frac{2}{3}T \right\rceil - \mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}]\right\}, \tag{A.2}$$

where Equation (A.1) follows from observing that we get reward from arm 2 only if we pull it more than $\lfloor \frac{T}{3} \rfloor$ times and Equation (A.2) derives from observing that $T = \mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] + \mathbb{E}_{\boldsymbol{\mu}^A}[N_{2,T}]$. Now, consider the two cases:

Case (i) : $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] \geq \lceil \frac{2}{3}T \rceil$

$$J_{\boldsymbol{\mu}^A}(\pi, T) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}],$$

that is maximized by taking $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] = T$.

Case (ii) : $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] < \lceil \frac{2}{3}T \rceil$

$$J_{\boldsymbol{\mu}^A}(\pi, T) = \left\lceil \frac{2}{3}T \right\rceil - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}],$$

that is maximized by taking the minimum value of $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}]$ possible, that is $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] \geq$

$\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,\lfloor \frac{T}{3} \rfloor}] = n_1$. Putting all together, we have:

$$J_{\boldsymbol{\mu}^A}(\pi, T) \leq \max\left\{\frac{T}{2}, \left\lceil \frac{2}{3}T \right\rceil - \frac{n_1}{2}\right\} = \left\lceil \frac{2}{3}T \right\rceil - \frac{n_1}{2},$$

having observed that $n_1 \leq \lfloor \frac{T}{3} \rfloor$. Let us now focus on the regret:

$$R_{\boldsymbol{\mu}^A}(\pi, T) = J^*_{\boldsymbol{\mu}^A}(T) - J_{\boldsymbol{\mu}^A}(\pi, T) = \left\lceil \frac{2}{3}T \right\rceil - \left\lceil \frac{2}{3}T \right\rceil + \frac{n_1}{2} = \frac{n_1}{2}.$$

Consider now bandit $\boldsymbol{\mu}^B$, we have:

$$J_{\boldsymbol{\mu}^B}(\pi, T) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}^B}[N_{1,T}] \leq \frac{n_1}{2} + \left\lfloor \frac{T}{3} \right\rfloor,$$

having observed that $\mathbb{E}_{\boldsymbol{\mu}^B}[N_{1,T}] = n_1 + \mathbb{E}_{\boldsymbol{\mu}^B}[N_{1,T}] - \mathbb{E}_{\boldsymbol{\mu}^B}[N_{1,\lfloor \frac{T}{3} \rfloor}] \leq n_1 + \lceil \frac{2}{3}T \rceil$. Let us now compute the regret:

$$R_{\boldsymbol{\mu}^B}(\pi, T) = J^*_{\boldsymbol{\mu}^B}(T) - J_{\boldsymbol{\mu}^B}(\pi, T) = \frac{T}{2} - \frac{n_1}{2} - \left\lfloor \frac{T}{3} \right\rfloor = \left\lceil \frac{T}{6} \right\rceil - \frac{n_1}{2}.$$

Finally, the worst-case regret can be lower bounded as follows:

$$\begin{aligned}
\inf_{\pi} \sup_{\boldsymbol{\mu}} R_{\boldsymbol{\mu}}(\pi, T) &\geq \inf_{\pi} \max\left\{R_{\boldsymbol{\mu}^A}(\pi, T), R_{\boldsymbol{\mu}^B}(\pi, T)\right\} \\
&\geq \inf_{n_1 \in \left[0, \lfloor \frac{T}{3} \rfloor\right]} \max\left\{\frac{n_1}{2}, \left\lceil \frac{T}{6} \right\rceil - \frac{n_1}{2}\right\} \\
&\geq \frac{1}{2}\left\lceil \frac{T}{6} \right\rceil \\
&\geq \left\lfloor \frac{T}{12} \right\rfloor,
\end{aligned}$$

having minimized over n_1. □

Theorem 4.2 (Non-Learnability). *There exists a 2-armed non-decreasing (non-concave) deterministic rested bandit with $\gamma_i(n) \leq \gamma_{\max} < 1$ for all $i \in [K]$ and $n \in \mathbb{N}$, such that any learning policy π suffers regret:*

$$R_{\boldsymbol{\mu}}(\pi, T) \geq \left\lfloor \frac{\gamma_{\max}}{12} T \right\rfloor.$$

Proof. It is sufficient to rescale the mean function of the proof of Lemma A.1 by the quantity $\gamma_{\max}$. □

A.1.1. Rested Upper Bound Derivation (Section 4.1.1)

Lemma 4.2 (Growth Bound). *Under Assumptions 2.1 and 2.2, for every $i \in [K]$, $k, k' \in \mathbb{N}$ with $k' < k$, for both restless and rested bandits, it holds that:*

$$\gamma_i(k) \leq \frac{\mu_i(k) - \mu_i(k')}{k - k'}.$$

Proof. Using Assumption 2.2, we have:

$$\gamma_i(k) = \frac{1}{k-k'}\sum_{l=k'}^{k-1}\gamma_i(k) \leq \frac{1}{k-k'}\sum_{l=k'}^{k-1}\gamma_i(l) = \frac{1}{k-k'}\sum_{l=k'}^{k-1}(\mu_i(l+1) - \mu_i(l)) = \frac{\mu_i(k) - \mu_i(k')}{k-k'},$$

where the first inequality comes from the concavity of the reward function, and the second equality from the definition of increment (2.1). □

Lemma 4.1 (Deterministic Rested UB). *For every arm $i \in [K]$ and every round $t \in [T]$:*

$$B_i(t) \equiv \overline{\mu}_i^{\texttt{R-ed}}(t) \geq \mu_i(t).$$

Moreover if $N_{i,t-1} \geq 2$ it holds that:

$$\overline{\mu}_i^{\texttt{R-ed}}(t) - \mu_i(N_{i,t}) \leq (t - N_{i,t-1})\gamma_i(N_{i,t-1} - 1).$$

Proof. Let us consider the following derivation:

$$\mu_i(t) = \mu_i(N_{i,t-1}) + \sum_{n=N_{i,t-1}}^{t-1}\gamma_i(n) \leq \mu_i(N_{i,t-1}) + (t - N_{i,t-1})\gamma_i(N_{i,t-1} - 1) =: \overline{\mu}_i^{\texttt{R-ed}}(t),$$

where the inequality holds thanks to Assumption 2.2, having observed that $\sum_{n=N_{i,t-1}}^{t-1}\gamma_i(n) \leq (t - N_{i,t-1})\gamma_i(N_{i,t-1}) \leq (t - N_{i,t-1})\gamma_i(N_{i,t-1} - 1)$.

For the bias bound, when $N_{i,t-1} \geq 2$, we consider the following derivation:

$$\begin{aligned}&\overline{\mu}_i^{\texttt{R-ed}}(t) - \mu_i(N_{i,t}) = \\ &\quad = \mu_i(N_{i,t-1}) + (t - N_{i,t-1})\gamma_i(N_{i,t-1} - 1) - \mu_i(N_{i,t}) \leq (t - N_{i,t-1})\gamma_i(N_{i,t-1} - 1).\end{aligned}$$

having observed that $\mu_i(N_{i,t-1}) \leq \mu_i(N_{i,t})$ by Assumption 2.1. □

Lemma 4.3 (Stochastic Rested UB). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widetilde{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(l) + (t-l) \frac{\mu_i(l) - \mu_i(l-h)}{h} \right),$$

otherwise if $h = 0$, we set $\widetilde{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq +\infty$. Then, $\widetilde{\mu}_i^{\texttt{R-ed},h}(t) \geq \mu_i(t)$.

Moreover, if $N_{i,t-1} \geq 2$, it holds that:

$$\widetilde{\mu}_i^{\texttt{R-ed},h}(t) - \mu_i(N_{i,t}) \leq \frac{1}{2}(2t - 2N_{i,t-1} + h - 1)\gamma_i(N_{i,t-1} - 2h + 1).$$

Proof. For every $l \in \{2, \dots, N_{i,t-1}\}$:

$$\begin{aligned} \mu_i(t) &= \mu_i(l) + \sum_{j=l}^{t-1} \gamma_i(j) \\ &\leq \mu_i(l) + (t-l)\gamma_i(l-1) && \text{(A.3)} \\ &\leq \mu_i(l) + (t-l)\frac{\mu_i(l) - \mu_i(l-h)}{h}, && \text{(A.4)} \end{aligned}$$

where line (A.3) follows from Assumption 2.2, line (A.4) is obtained from Lemma 4.2.

By averaging over the most recent $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$ pulls, we obtain:

$$\mu_i(t) \leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(l) + (t-l) \frac{\mu_i(l) - \mu_i(l-h)}{h} \right) =: \widetilde{\mu}_i^{\texttt{R-ed},h}(t).$$

For the bias bound, when $N_{i,t-1} \geq 2$, we have:

$$\widetilde{\mu}_i^{\texttt{R-ed},h}(t) - \mu_i(N_{i,t}) = \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(l) + (t-l)\frac{\mu_i(l) - \mu_i(l-h)}{h} \right) - \mu_i(N_{i,t}) \quad \text{(A.5)}$$

$$\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l)\frac{\mu_i(l) - \mu_i(l-h)}{h}$$

$$= \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l)\frac{1}{h} \sum_{j=l-h}^{l-1} \gamma_j(l)$$

$$\leq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} (t-l)\gamma_i(l-h) \quad \text{(A.6)}$$

$$\leq \frac{1}{2}(2t - 2N_{i,t-1} + h - 1)\gamma_i(N_{i,t-1} - 2h + 1). \quad \text{(A.7)}$$

where line (A.5) follows from Assumption 2.1 applied as $\mu_i(l) \leq \mu_i(N_{i,t})$, line (A.6) follows from Assumption 2.2 and bounding $\frac{1}{h}\sum_{j=l-h}^{l-1} \gamma_j(l) \leq \gamma_i(l-h)$ and line (A.7) is derived still from Assumption 2.2, $\gamma_i(l-h) \leq \gamma_i(N_{i,t-1} - 2h + 1)$ and computing the summation. □

Lemma 4.4 (Rested UB Concentration). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widehat{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(R_{t_{i,l}} + (t-l)\frac{R_{t_{i,l}} - R_{t_{i,l-h}}}{h} \right),$$

$$\beta_i^{\texttt{R-ed},h}(t,\delta) \coloneqq \sigma(t - N_{i,t-1} + h - 1)\sqrt{\frac{10\log\frac{1}{\delta}}{h^3}},$$

otherwise if $h = 0$, we set $\widehat{\mu}_i^{\texttt{R-ed},h}(t) \coloneqq +\infty$ and $\beta_i^{\texttt{R-ed},h}(t,\delta) \coloneqq +\infty$. 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:

$$\Pr\left(\left| \widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) \right| > \beta_i^{\texttt{R-ed},h_{i,t}}(t,\delta_t) \right) \leq 2t^{1-\alpha}.$$

Proof. First of all, we observe under the event $\{h_{i,t} = 0\}$ that $\widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) = \widetilde{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) = \beta_i^{\texttt{R-ed},h_{i,t}}(t,\delta_t) = +\infty$. By convening that $(+\infty)-(+\infty) = 0$, we have that $0 > \beta_i^{\texttt{R-ed},h_{i,t}}(t,\delta_t)$ is not satisfied. Thus, we perform the analysis under the event $\{h_{i,t} \geq 1\}$. We first get

rid of the dependence on the random number of pulls $N_{i,t-1}$:

$$\begin{aligned}
&\Pr\left(\left|\widehat{\mu}_i^{\texttt{R-ed},h_{i,t}}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h_{i,t}}(t)\right| > \beta_i^{\texttt{R-ed},h_{i,t}}(t,\delta_t)\right) \\
&\quad = \Pr\left(\left|\widehat{\mu}_i^{\texttt{R-ed},h(N_{i,t-1})}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h(N_{i,t-1})}(t)\right| > \beta_i^{\texttt{R-ed},h(N_{i,t-1})}(t,\delta_t)\right) \\
&\quad \leq \Pr\left(\exists n \in [t-1] \text{ s.t. } h(n) \geq 1 \,:\, \left|\widehat{\mu}_i^{\texttt{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h(n)}(t)\right| > \beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right) \\
&\quad \leq \sum_{n\in\{0,\dots,t-1\}\,:\,h(n)\geq 1} \Pr\left(\left|\widehat{\mu}_i^{\texttt{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h(n)}(t)\right| > \beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right),
\end{aligned}$$

where the first equation derives from the definition of $h_{i,t} = h(N_{i,t-1})$ and the last line follows from a union bound over the possible values of $N_{i,t-1}$. Now, having fixed the value of n, we rewrite the quantity to be bounded:

$$\begin{aligned}
h(n)\left(\widehat{\mu}_i^{\texttt{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h(n)}(t)\right) &= \sum_{l=n-h(n)+1}^{n}\left(X_l + (t-l)\frac{X_l - X_{l-h(n)}}{h(n)}\right) \\
&= \sum_{l=n-h(n)+1}^{n}\left(1+\frac{t-l}{h(n)}\right)X_l - \sum_{l=n-h(n)+1}^{n}\frac{t-l}{h(n)}\cdot X_{l-h(n)},
\end{aligned}$$

where $X_l \coloneqq R_{t_{i,l}} - \mu_i(l)$. It is worth noting that we can index X_l with the number of pulls l only as the distribution of $R_{t_{i,l}}$ is fully determined by l and n (that are non-random quantities now) and, consequently, all variables X_l and $X_{l-h(n)}$ are independent.

Now we apply Azuma-Hoëffding's inequality of Lemma A.1 for weighted sums of subgaussian martingale difference sequences. To this purpose, we compute the sum of the square weights:

$$\begin{aligned}
&\sum_{l=n-h(n)+1}^{n}\left(1+\frac{t-l}{h(n)}\right)^2 + \sum_{l=n-h(n)+1}^{n}\left(\frac{t-l}{h(n)}\right)^2 \\
&\quad \leq h(n)\left(1+\frac{t-n+h(n)-1}{h(n)}\right)^2 + h(n)\left(\frac{t-n+h(n)-1}{h(n)}\right)^2 \qquad \text{(A.8)} \\
&\quad \leq \frac{5(t-n+h(n)-1)^2}{h(n)}, \qquad \text{(A.9)}
\end{aligned}$$

where line (A.8) follows from bounding $t - l \leq t - n + h(n) - 1$ and line (A.9) from

observing that $\frac{t-n+h(n)-1}{h(n)} \geq 1$. Thus, we have:

$$\begin{aligned}
&\Pr\left(\left|\widehat{\mu}_i^{\texttt{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h(n)}(t)\right| > \beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right) \\
&\quad \leq \Pr\left(\left|\sum_{l=n-h(n)+1}^{n}\left(1+\frac{t-l}{h(n)}\right)X_l - \sum_{l=n-h(n)+1}^{n}\frac{t-l}{h(n)}\cdot X_{l-h(n)}\right| > h(n)\beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right) \\
&\quad \leq 2\exp\left(-\frac{\left(h(n)\beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right)^2}{2\sigma^2\left(\frac{5(t-n+h(n)-1)^2}{h(n)}\right)}\right) = 2\delta_t.
\end{aligned}$$

By substituting the result into the previously found

$$\sum_{n\in\{0,\dots,t-1\}\,:\,h(n)\geq 1} \Pr\left(\left|\widehat{\mu}_i^{\texttt{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-ed},h(n)}(t)\right| > \beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right),$$

and recalling the value of δ_t, we obtain:

$$\sum_{n\in\{0,\dots,t-1\}\,:\,h(n)\geq 1} 2\delta_t \leq \sum_{n=0}^{t-1} 2\delta_t = \sum_{n=0}^{t-1} 2t^{-\alpha} \leq 2t^{1-\alpha}.$$

□

A.2. Proofs Restless Setting (Section 4.2)

Theorem 4.5 (Greedy Policy Optimality [52]). *Let $\pi^g_{\boldsymbol{\mu}}$ be the* oracle greedy *policy:*

$$\pi^g_{\boldsymbol{\mu}}(t) \in \operatorname*{argmax}_{i\in[K]}\{\mu_i(t)\}, \quad \forall t \in [T].$$

Then, $\pi^g_{\boldsymbol{\mu}}$ is optimal for the restless non-decreasing bandits (i.e., under Assumption 2.1).

Proof. Trivially follows from the fact that the greedy policy at each round t is selecting the largest expected reward, therefore any optimal policy other than the greedy one should select a larger expected reward at least for a single round t', which is in contradiction with the definition of greedy policy. □

A.2.1. Restless Upper Bound Derivation (Section 4.2.1)

Lemma 4.6 (Deterministic Restless UB). *For every arm $i \in [K]$ and every round $t \in [T]$:*

$$B_i(t) \equiv \overline{\mu}_i^{\texttt{R-less}}(t) \geq \mu_i(t).$$

Moreover, if $N_{i,t-1} \geq 2$, it holds that:

$$\overline{\mu}_i^{\texttt{R-less}}(t) - \mu_i(t) \leq \left(t - t_{i,N_{i,t-1}}\right) \gamma_i(t_{i,N_{i,t-1}-1}).$$

Proof. Let us consider the following derivation:

$$\begin{aligned}
\mu_i(t) &= \mu_i(t_{i,N_{i,t-1}}) + \sum_{l=t_{i,N_{i,t-1}}}^{t-1} \gamma_i(l) \\
&\leq \mu_i(t_{i,N_{i,t-1}}) + (t - t_{i,N_{i,t-1}})\gamma_i(t_{i,N_{i,t-1}}) && \text{(A.10)} \\
&\leq \mu_i(t_{i,N_{i,t-1}}) + (t - t_{i,N_{i,t-1}})\frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}} =: \overline{\mu}_i^{\texttt{R-less}}(t), && \text{(A.11)}
\end{aligned}$$

where line (A.10) follows from Assumption 2.2 and line (A.11) from Lemma 4.2.

Moreover, if $N_{i,t-1} \geq 2$, we have:

$$\begin{aligned}
\overline{\mu}_i^{\texttt{R-less}}(t) - \mu_i(t) &= (t - t_{i,N_{i,t-1}}) \frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}} + \underbrace{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t)}_{\leq 0} \\
&\leq (t - t_{i,N_{i,t-1}}) \frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}} \\
&= \frac{t - t_{i,N_{i,t-1}}}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}} \sum_{l=t_{i,N_{i,t-1}-1}}^{t_{i,N_{i,t-1}}-1} \gamma_i(l), \\
&\leq \left(t - t_{i,N_{i,t-1}}\right) \gamma_i(t_{i,N_{i,t-1}-1}),
\end{aligned}$$

where in the last line we employed Assumption 2.2, noting that:

$$\frac{1}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}} \sum_{l=t_{i,N_{i,t-1}-1}}^{t_{i,N_{i,t-1}}-1} \gamma_i(l) \leq \gamma_i(t_{i,N_{i,t-1}-1})$$

□

Lemma 4.7 (Stochastic Restless UB). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widetilde{\mu}_i^{\texttt{R-less},h}(t) :- \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(\mu_i(t_{i,l}) + (t-l) \frac{\mu_i(t_{i,l}) - \mu_i(t_{i,l-h})}{h} \right),$$

otherwise if $h = 0$, we set $\widetilde{\mu}_i^{\texttt{R-less},h}(t) := +\infty$. Then, $\widetilde{\mu}_i^{\texttt{R-less},h}(t) \geq \mu_i(t_{i,N_{i,t-1}})$.

Moreover, if $N_{i,t-1} \geq 2$ it holds that:

$$\widetilde{\mu}_i^{\texttt{R-less},h}(t) - \mu_i(t) \leq \frac{(2t - 2N_{i,t-1} + h - 1)(t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-2h+1})}{2h} \gamma_i(t_{i,N_{i,t-1}-2h+1}).$$

Proof. Let us start by observing the following equality holding for every $l \in \{2, \dots, N_{i,t-1}\}$:

$$\mu_i(t) = \mu_i(t_{i,l}) + \sum_{j=t_{i,l}}^{t-1} \gamma_i(j).$$

By averaging over a window of length h, we obtain:

$$\begin{aligned}
\mu_i(t) &= \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l}) + \sum_{j=t_{i,l}}^{t-1}\gamma_i(j)\right) \\
&\leq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l}) + (t-t_{i,l})\gamma_i(t_{i,l}-1)\right) && \text{(A.12)} \\
&\leq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l}) + \frac{t-t_{i,l}}{t_{i,l}-t_{i,l-h}}\sum_{j=t_{i,l-h}}^{t_{i,l}-1}\gamma_i(j)\right) && \text{(A.13)} \\
&= \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l}) + (t-t_{i,l})\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{t_{i,l}-t_{i,l-h}}\right) \\
&\leq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l}) + (t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h}\right) =: \widetilde{\mu}_i^{\texttt{R-less},h}(t), && \text{(A.14)}
\end{aligned}$$

where lines (A.12) and (A.13) follow from Assumption 2.2, and line (A.14) is obtained from observing that $t_{i,l} \geq l$ and $t_{i,l} - t_{i,l-h} \geq h$.

Concerning the bias, when $N_{i,t-1} \geq 2$, we have:

$$\begin{aligned}
&\widetilde{\mu}_i^{\texttt{R-less},h}(t) - \mu_i(t) = \\
&\quad= \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l}) + (t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h}\right) - \mu_i(t) \\
&\quad\leq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h} && \text{(A.15)} \\
&\quad= \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{t_{i,l}-t_{i,l-h}}\cdot\frac{t_{i,l}-t_{i,l-h}}{h} \\
&\quad\leq \frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\gamma_i(t_{i,l-h})\cdot\frac{t_{i,l}-t_{i,l-h}}{h} && \text{(A.16)} \\
&\quad\leq \frac{t_{i,N_{i,t-1}}-t_{i,N_{i,t-1}-2h+1}}{h^2}\gamma_i(t_{i,N_{i,t-1}-2h+1})\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l) && \text{(A.17)} \\
&\quad= \frac{(2t-2N_{i,t-1}+h-1)(t_{i,N_{i,t-1}}-t_{i,N_{i,t-1}-2h+1})}{2h}\gamma_i(t_{i,N_{i,t-1}-2h+1}), && \text{(A.18)}
\end{aligned}$$

where line (A.15) follows from observing that $\mu_i(t_{i,l}) \leq \mu_i(t)$, line (A.16) derives from Assumption 2.2 and bounding $\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{t_{i,l}-t_{i,l-h}} \leq \gamma_i(t_{i,l-h})$, line (A.17) is obtained by bounding

$t_{i,l} - t_{i,l-h} \leq t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-2h+1}$ and $\gamma_i(t_{i,l-h}) \leq \gamma_i(t_{i,N_{i,t-1}-2h+1})$, and line (A.18) follows from computing the summation. □

Lemma 4.8 (Restless UB Concentration). *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leq h \leq \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

$$\widehat{\mu}_i^{\texttt{R-less},h}(t) \coloneqq \frac{1}{h} \sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}} \left(R_{t_{i,l}} + (t-l)\frac{R_{t_{i,l}} - R_{t_{i,l-h}}}{h} \right),$$

$$\beta_i^{\texttt{R-less},h}(t,\delta) \coloneqq \sigma(t - N_{i,t-1} + h - 1)\sqrt{\frac{10 \log \frac{1}{\delta}}{h^3}},$$

otherwise if $h = 0$, we set $\widehat{\mu}_i^{\texttt{R-less},h}(t) \coloneqq +\infty$ and $\beta_i^{\texttt{R-less},h}(t,\delta) \coloneqq +\infty$. 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:

$$\Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) - \widetilde{\mu}_i^{\texttt{R-less},h_{i,t}}(t) \right| > \beta_i^{\texttt{R-less},h_{i,t}}(t,\delta_t) \right) \leq 2t^{1-\alpha}.$$

Proof. Under the event $\{h_{i,t} = 0\}$, we have that $\widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) = \widetilde{\mu}_i^{\texttt{R-less},h_{i,t}}(t) = \beta_i^{\texttt{R-less},h_{i,t}}(t,\delta) = +\infty$ and, under the convention $(+\infty) - (+\infty) = 0$ the event $0 > \beta_i^{\texttt{R-less},h}(t,\delta)$ does not hold. Therefore, we conduct the proof under the event $\{h_{i,t} \geq 1\}$. Hence:

$$\Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) - \widetilde{\mu}_i^{\texttt{R-less},h_{i,t}}(t) \right| > \beta_i^{\texttt{R-less},h_{i,t}}(t,\delta_t) \right) = \tag{A.19}$$

$$= \Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h(N_{i,t-1})}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(N_{i,t-1})}(t) \right| > \beta_i^{\texttt{R-less},h(N_{i,t-1})}(t,\delta_t) \right) \tag{A.20}$$

$$\leq \Pr\left(\exists n \in [t-1] \text{ s.t. } h(n) \geq 1 : \left| \widehat{\mu}_i^{\texttt{R-less},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(n)}(t) \right| > \beta_i^{\texttt{R-less},h(n)}(t,\delta_t) \right)$$

$$\leq \sum_{n \in \{0,\dots,t-1\} : h(n) \geq 1} \Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(n)}(t) \right| > \beta_i^{\texttt{R-less},h(n)}(t,\delta_t) \right), \tag{A.21}$$

where line (A.20) follows from the definition of $h_{i,t} = h(N_{i,t-1})$, and line (A.21) derives from a union bound over n.

Differently from the rested case, in which the distribution of all random variable involved is fully determined having fixed $N_{i,t-1}$, in the restless case this is no longer true. Indeed, the distribution of the rewards does not depend on the number of pulls, but on the round in which the arm was pulled. Thus, we need a more articulated argument. We start

rewriting the estimator with a summation over rounds:

$$\begin{aligned} & h(n)\left(\widehat{\mu}_i^{\texttt{R-less},h(n)}(t)-\widetilde{\mu}_i^{\texttt{R-less},h(n)}(t)\right) \\ & \quad=\sum_{l=n-h(n)+1}^{n}\left(X_{t_{i,l}}+(t-l)\frac{X_{t_{i,l}}-X_{t_{i,l-h(n)}}}{h(n)}\right) \\ & \quad=\sum_{s=1}^{t-1}\epsilon_s Y_s X_s, \end{aligned} \tag{A.22}$$

where:

$$\begin{aligned} &\epsilon_s=\mathbb{1}\{I_s=i\}, \\ &X_s=R_s-\mu_i(s) \\ &Y_s=\left(\mathbb{1}\left\{N_{i,s}\in\{n-h(n)+1,\dots,n\}\right\}\left(1+\frac{t-N_{i,s}}{h(n)}\right)\right. \\ &\qquad\left.-\mathbb{1}\left\{N_{i,s}\in\{n-2h(n)+1,\dots,n-h(n)\}\right\}\frac{t-N_{i,s}-h(n)}{h(n)}\right), \end{aligned}$$

The rationale behind this decomposition is to use random variable ϵ_s to select the pulls of arm i, Y_s to define the quantity by which X_s is multiplied. In particular, if the pull belongs to the set of the most recent $h(n)$ pulls, i.e., $N_{i,s}\in\{n-h(n)+1,\dots,n\}$, we multiply X_s by the constant $1+\frac{t-N_{i,s}}{h(n)}$. Instead, if the pull belongs to less recent $h(n)$ pulls, i.e., $N_{i,s}\in\{n-2h(n)+1,\dots,n-h(n)\}$, we multiply X_s by $\frac{t-N_{i,s}-h(n)}{h(n)}$. Now, we define the sequence of random times at which arm i was pulled for the j-th time:

$$t_{i,j}\coloneqq\min_{t\in[T]}\{N_{i,t}=j\},\quad j\in[n],$$

and we introduce the random variables $\widetilde{X}_j\coloneqq X_{t_{i,j}}$ and $\widetilde{Y}_j\coloneqq Y_{t_{i,j}}$. To prove that $\widetilde{Y}_j\widetilde{X}_j$ is a martingale difference sequence w.r.t.to the filtration it generates, we apply a Doob's *optional skipping* argument [14, 20].

We introduce the filtration $\mathcal{F}_{\tau-1}=\sigma\left(I_1,R_1,\dots,I_{\tau-1},R_{\tau-1},I_\tau\right)$ and we need to show that: (i) $Z_\tau=\sum_{s=1}^{\tau}\epsilon_s Y_s X_s$ is a martingale, and (ii) $\{t_{i,j}=\tau\}\in\mathcal{F}_{\tau-1}$ for $\tau\in[t-1]$.

Concerning (i), we have:

$$\mathbb{E}\left[Z_\tau|\mathcal{F}_{\tau-1}\right]=Z_{\tau-1}+\epsilon_\tau Y_\tau\mathbb{E}\left[X_\tau|\mathcal{F}_{\tau-1}\right]=Z_{\tau-1},$$

since $\epsilon_\tau Y_\tau$ is fully determined by $\mathcal{F}_{\tau-1}$ and either $\epsilon_\tau=0$ or $I_\tau=i$, thus, $\epsilon_\tau\mathbb{E}\left[X_\tau|\mathcal{F}_{\tau-1}\right]=\epsilon_\tau\mathbb{E}\left[R_\tau-\mu_i(\tau)|\mathcal{F}_{\tau-1}\right]=0$.

Concerning (ii), $\{t_{i,j} = \tau\} \in \mathcal{F}_{\tau-1}$ is trivially verified.

We recall that, since $\widetilde{Y}_j = Y_{t_{i,j}}$ we have that $N_{i,t_{i,j}} = j$:

$$\widetilde{Y}_j = \left(\mathbb{1}\left\{ j \in \{n - h(n) + 1, \dots, n\} \right\} \left(1 + \frac{t-j}{h(n)} \right) \right.$$
$$\left. - \mathbb{1}\left\{ j \in \{n - 2h(n) + 1, \dots, n - h(n)\} \right\} \frac{t - j - h(n)}{h(n)} \right).$$

From which, by substituting into Equation (A.22) and properly solving the indicator functions, we have:

$$\sum_{j=1}^{n} \widetilde{X}_j \widetilde{Y}_j = \sum_{j=n-h(n)+1}^{n} \left(1 + \frac{t-j}{h(n)} \right) \widetilde{X}_j - \sum_{j=n-2h(n)+1}^{n-h(n)} \frac{t-j}{h(n)} \cdot \widetilde{X}_j.$$

We compute the square of the weights and apply a derivation similar to that of Lemma 4.4:

$$\sum_{j=n-h(n)+1}^{n} \left(1 + \frac{t-j}{h(n)} \right)^2 + \sum_{j=n-2h(n)+1}^{n-h(n)} \left(\frac{t-j}{h(n)} \right)^2 \leq \frac{5(t - n + h(n) - 1)^2}{h(n)}.$$

Thus, we can now apply Azuma-Höeffding's inequality (Lemma A.1):

$$\begin{aligned}
&\Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(n)}(t) \right| > \beta_i^{\texttt{R-less},h(n)}(t, \delta_t) \right) \\
&\quad = \Pr\left(\left| \sum_{s=1}^{t} \epsilon_s X_s Y_s \right| > h(n) \beta_i^{\texttt{R-less},h(n)}(t, \delta_t) \right) \\
&\quad = \Pr\left(\left| \sum_{j=1}^{n} \widetilde{X}_j \widetilde{Y}_j \right| > h(n) \beta_i^{\texttt{R-less},h(n)}(t, \delta_t) \right) \\
&\quad \leq 2 \exp\left(- \frac{\left(h(n) \beta_i^{\texttt{R-ed},h(n)}(t, \delta_t) \right)^2}{2\sigma^2 \left(\frac{5(t-n+h(n)-1)^2}{h(n)} \right)} \right) = 2\delta_t.
\end{aligned}$$

By replacing into Equation (A.21) and summing over n, we obtain:

$$\sum_{n \in \{0,\dots,t-1\} \,:\, h(n) \geq 1} 2\delta_t \leq \sum_{n=0}^{t-1} 2\delta_t = 2t^{1-\alpha}.$$

□

A.3. Technical Lemmas

Lemma A.2. *Let $M \geq 3$, and let $f : \mathbb{N} \to \mathbb{R}$, and $\beta \in (0, 1)$. Then it holds that:*

$$\sum_{j=3}^{M} f(\lfloor \beta j \rfloor) \leq \left\lceil \frac{1}{\beta} \right\rceil \sum_{l=\lfloor 3\beta \rfloor}^{\lfloor \beta M \rfloor} f(l).$$

Proof. We simply observe that the minimum value of $\lfloor \beta j \rfloor$ is $\lfloor 3\beta \rfloor$ and its maximum value is $\lfloor \beta M \rfloor$. Each element $\lfloor \beta j \rfloor$ changes value at least one time every $\left\lceil \frac{1}{\beta} \right\rceil$ times. □

Lemma A.3. *Under Assumption 2.2, it holds that:*

$$\max_{\substack{(N_{i,T})_{i\in[K]} \\ N_{i,T}\geq 0, \sum_{i\in[K]} N_{i,T}=T}} \sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}(N_{i,T}, q) \leq K \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil \frac{T}{K} \right\rceil, q \right).$$

Proof. We first claim that there exists an optimal assignment of $N^*_{i,T}$ are such that $|N^*_{i,T} - N^*_{i',T}| \leq 1$ for all $i, i' \in [K]$. By contradiction, suppose that the only optimal assignments are such that there exists a pair $i_1, i_2 \in [K]$ such that $\Delta := N^*_{i_2,T} - N^*_{i_1,T} > 1$. In such a case, we have:

$$\begin{aligned}
&\Upsilon_{\boldsymbol{\mu}}\left(N^*_{i_1,T}, q\right) + \Upsilon_{\boldsymbol{\mu}}\left(N^*_{i_2,T}, q\right) \\
&\quad = 2\Upsilon_{\boldsymbol{\mu}}\left(N^*_{i_1,T}, q\right) + \sum_{j=1}^{\Delta} \gamma_{i^*}(N^*_{i_1,T} + l - 1) \\
&\quad \leq 2\Upsilon_{\boldsymbol{\mu}}\left(N^*_{i_1,T}, q\right) + \sum_{j=0}^{\lceil \Delta/2 \rceil} \gamma_{i^*}(N^*_{i_1,T} + l - 1) + \sum_{j=1}^{\lfloor \Delta/2 \rfloor} \gamma_{i^*}(N^*_{i_1,T} + l - 1) \\
&\quad = \Upsilon_{\boldsymbol{\mu}}\left(N^*_{i_1,T} + \lceil \Delta/2 \rceil, q\right) + \Upsilon_{\boldsymbol{\mu}}\left(N^*_{i_1,T} + \lfloor \Delta/2 \rfloor, q\right).
\end{aligned}$$

where the inequality follows from Assumption 2.2. By redefining $\widetilde{N}^*_{i_1,T} := N^*_{i_1,T} + \lfloor \Delta/2 \rfloor$ and $\widetilde{N}^*_{i_2,T} := N^*_{i_1,T} + \lceil \Delta/2 \rceil$, we have that $\widetilde{N}^*_{i_1,T} + \widetilde{N}^*_{i_2,T} = N^*_{i_1,T} + N^*_{i_2,T}$ and $|\widetilde{N}^*_{i_1,T} - \widetilde{N}^*_{i_2,T}| \leq 1$. Thus, we have found a better solution to the optimization problem, contradicting the hypothesis. Since the optimal assignment fulfills $|N^*_{i,T} - N^*_{i',T}| \leq 1$, it must be that $N^*_{i,T} \leq \left\lceil \frac{T}{K} \right\rceil$ for all $i \in [K]$. □

Lemma A.4. *Let $a, b \in \mathbb{N}$ and let $f : [a, b] \to \mathbb{R}$. If f is monotonically non-decreasing*

function, then:

$$\sum_{n=a}^{b} f(n) \leq \int_{x=a}^{b} f(x)\mathrm{d}x + f(b) \leq \int_{x=a}^{b+1} f(x).$$

If f is monotonically non-increasing, then:

$$\sum_{n=a}^{b} f(n) \leq f(a) + \int_{x=a}^{b} f(x)\mathrm{d}x \leq \int_{x=a-1}^{b} f(x)\mathrm{d}x.$$

Proof. Let us consider the intervals $I_i = [x_{i-1}, x_i]$ with $x_0 = a$ and $x_i = x_{i-1} + 1$ for $i \in [b-a]$. If f is monotonically non-decreasing, we have that for all $i \in [b-a]$ and $x \in I_i$ it holds that $f(x) \geq f(x_{i-1})$ and consequently $\int_{I_i} f(x)\mathrm{d}x \geq f(x_{i-1})\mathrm{vol}(I_i) = f(x_{i-1})$. Thus:

$$\sum_{n=a}^{b} f(n) = \sum_{i=1}^{b-a} f(x_{i-1}) + f(b) \leq \sum_{i=1}^{b-a} \int_{I_i} f(x)\mathrm{d}x + f(b) = \int_{x=a}^{b} f(x)\mathrm{d}x + f(b).$$

Recalling that $f(b) \leq \int_{x=b}^{b+1} f(x)\mathrm{d}x$, we get the second inequality. Conversely, if f is monotonically non-increasing, then for all $i \in [b-a]$ and $x \in I_i$, it holds that $f(x) \geq f(x_i)$ and consequently $\int_{I_i} f(x)\mathrm{d}x \geq f(x_i)$. Thus:

$$\sum_{n=a}^{b} f(n) = f(a) + \sum_{i=1}^{b-a} f(x_i) \leq f(a) + \sum_{i=1}^{b-a} \int_{I_i} f(x)\mathrm{d}x = f(a) + \int_{x=a}^{b} f(x)\mathrm{d}x.$$

Recalling that $f(a) \leq \int_{x=a-1}^{a} f(x)\mathrm{d}x$, we get the second inequality. □

Theorem A.1 (Höeffding-Azuma's inequality for weighted martingales)**.** *Let $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration and $X_1, \ldots, X_n$ be real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$ (i.e.,a martingale difference sequence), and $\mathbb{E}[\exp(\lambda X_t)|\mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ for any $\lambda > 0$ (i.e.,σ^2-subgaussian). Let $\alpha_1, \ldots, \alpha_n$ be non-negative real numbers. Then, for every $\kappa \geq 0$ it holds that:*

$$\Pr\left(\left|\sum_{t=1}^{n} \alpha_t X_t\right| > \kappa\right) \leq 2\exp\left(-\frac{\kappa^2}{2\sigma^2 \sum_{t=1}^{n} \alpha_i^2}\right).$$

Proof. It is a straightforward extension of Azuma-Höeffding inequality for subgaussian

random variables. We apply the Chernoff's method for some $s > 0$:

$$\Pr\left(\sum_{t=1}^{n} \alpha_t X_t > \kappa\right) = \Pr\left(e^{s\sum_{t=1}^{n}\alpha_t X_t} > e^{s\kappa}\right) \leq \frac{\mathbb{E}\left[e^{s\sum_{t=1}^{n}\alpha_t X_t}\right]}{e^{s\kappa}},$$

where the last inequality follows from the application of Markov's inequality. We use the martingale property to deal with the expectation. By the law of total expectation, we have:

$$\mathbb{E}\left[e^{s\sum_{t=1}^{n}\alpha_t X_t}\right] = \mathbb{E}\left[e^{s\sum_{t=1}^{n-1}\alpha_t X_t}\mathbb{E}\left[e^{s\alpha_n X_n}|\mathcal{F}_{t-1}\right]\right].$$

Using now the subgaussian property, we have:

$$\mathbb{E}\left[e^{s\alpha_n X_n}|\mathcal{F}_{t-1}\right] \leq \exp\left(\frac{s^2\alpha_n^2\sigma^2}{2}\right).$$

An inductive argument, leads to:

$$\mathbb{E}\left[e^{s\sum_{t=1}^{n}\alpha_t X_t}\right] \leq \exp\left(\frac{s^2\sigma^2}{2}\sum_{t=1}^{n}\alpha_n^2\right).$$

Thus, minimizing w.r.t.$s > 0$, we have:

$$\Pr\left(\sum_{t=1}^{n} \alpha_t X_t > \kappa\right) \leq \min_{s\geq 0} \exp\left(\frac{s^2\sigma^2}{2}\sum_{t=1}^{n}\alpha_n^2 - s\kappa\right) = \exp\left(-\frac{\kappa^2}{2\sigma^2\sum_{t=1}^{n}\alpha_n^2}\right),$$

being the minimum attained by $s = \frac{\kappa}{\sigma^2\sum_{t=1}^{n}\alpha_n^2}$. The reverse inequality can be derived analogously. A union bound completes the proof. □

Lemma 4.5. *Let $\Upsilon_{\boldsymbol{\mu}}(M, q)$ be as defined in Equation (4.2) for some $q \in [0, 1]$. Then, for all $i \in [K]$ and $l \in \mathbb{N}$ the following statements hold:*

- *if $\gamma_i(l) \leq be^{-cl}$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}\left(b^q \frac{e^{-cq}}{cq}\right)$;*
- *if $\gamma_i(l) \leq bl^{-c}$ with $cq > 1$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}\left(\frac{b^q}{cq-1}\right)$;*
- *if $\gamma_i(l) \leq bl^{-c}$ with $cq = 1$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}\left(b^q \log M\right)$;*
- *if $\gamma_i(l) \leq bl^{-c}$ with $cq < 1$, then $\Upsilon_{\boldsymbol{\mu}}(M, q) \leq \mathcal{O}\left(b^q \frac{M^{1-cq}}{1-cq}\right)$.*

Proof. The proofs of all the statements are obtained by bounding the summation defining $\Upsilon_{\boldsymbol{\mu}}(T, q)$ with the corresponding integrals, as in Lemma A.4. Let us start with $\gamma_i(l) \leq$

be^{-cl}:

$$\Upsilon_{\boldsymbol{\mu}}(T,q) = \sum_{l=1}^{T} \gamma_i(l)^q \leq b^q e^{-cq} + \int_{x=1}^{T} b^q e^{-cqx} \mathrm{d}x \leq b^q e^{-cq} + \frac{b^q}{cq} e^{-cq} = \mathcal{O}\left(b^q \frac{e^{-cq}}{cq}\right).$$

We now move to $\gamma_i(l) \leq bl^{-c}$. If $cq < 1$, we have:

$$\Upsilon_{\boldsymbol{\mu}}(T,q) = \sum_{l=1}^{T} \gamma_i(l)^q \leq b^q + \int_{x=1}^{T} b^q x^{-cq} \mathrm{d}x = b^q + \frac{b^q}{cq-1} = \mathcal{O}\left(\frac{b^q}{cq-1}\right).$$

For $cq = 1$, we obtain:

$$\Upsilon_{\boldsymbol{\mu}}(T,q) = \sum_{l=1}^{T} \gamma_i(l)^q \leq b^q + \int_{x=1}^{T} \frac{b^q}{x} \mathrm{d}x = b^q + b^q \log T = \mathcal{O}\left(b^q \log T\right).$$

Finally, for $cq < 1$, we have:

$$\Upsilon_{\boldsymbol{\mu}}(T,q) = \sum_{l=1}^{T} \gamma_i(l)^q \leq b^q + \int_{x=1}^{T} b^q x^{-cq} \mathrm{d}x = b^q + b^q \frac{T^{1-cq}}{1-cq} = \mathcal{O}\left(b^q \frac{T^{1-cq}}{1-cq}\right).$$

The results of Table 4.1 are obtained by setting $b = 1$. □

B | Additional Results

B.1. Bounding the Cumulative Increment

Let us consider the case in which $\gamma_i(l) \leq l^{-c}$ for all $i \in [K]$ and $l \in [T]$. We bound the cumulative increment with the corresponding integral using Lemma A.4, depending on the value of cq:

$$\Upsilon_{\boldsymbol{\mu}}\left(\left\lceil \frac{T}{K} \right\rceil, q\right) = \sum_{l=1}^{\left\lceil \frac{T}{K} \right\rceil} \gamma_i(l)^q \leq 1 + \int_{x=1}^{\frac{T}{K}} x^{-cq} \mathrm{d}x \leq 1 + \begin{cases} \left(\frac{T}{K}\right)^{1-cq} \frac{1}{1-cq} & \text{if } cq < 1 \\ \log \frac{T}{K} & \text{if } cq = 1 \\ \frac{1}{cq-1} & \text{if } cq > 1 \end{cases}.$$

Thus, depending on the value of c, there will be different optimal values for q in the rested and restless cases that optimize the regret upper bound.

B.1.1. Rested Setting

Let us start with the rested case. From Theorem 4.3, we have:

$$R_{\boldsymbol{\mu}} \leq 2K + T^q K \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil \frac{T}{K} \right\rceil, q\right) \leq 2K + KT^q + K \begin{cases} \frac{T^{1-cq+q}}{K^{1-qc}(1-cq)} & \text{if } cq < 1 \\ T^q \log \frac{T}{K} & \text{if } cq = 1 \\ \frac{T^q}{cq-1} & \text{if } cq > 1 \end{cases}$$

$$\leq \mathcal{O}\left(K \begin{cases} \frac{T^{1-cq+q}}{K^{1-qc}(1-cq)} & \text{if } cq < 1 \\ T^q \log \frac{T}{K} & \text{if } cq = 1 \\ \frac{T^q}{\min\{1, cq-1\}} & \text{if } cq > 1 \end{cases} \right) \qquad \forall q \in [0,1],$$

where we have highlighted the dominant term. For the case $c \in (0,1)$ we consider the first case only and minimize over q:

$$R_{\boldsymbol{\mu}} \leq \mathcal{O}\left(K \min_{q \in [0,1]} \frac{T^{1-cq+q}}{K^{1-qc}(1-cq)} \right) = \mathcal{O}\left(T\right).$$

For the case $c = 1$, we still obtain $R_{\boldsymbol{\mu}} \leq \mathcal{O}(T)$. Instead, for $c \in (1, +\infty)$, we have the three cases:

$$R_{\boldsymbol{\mu}} \leq \mathcal{O}\left(K \min\begin{cases} K \min_{q\in[0,1/c)} \frac{T^{1-cq+q}}{K^{1-qc}(1-cq)} \\ T^{\frac{1}{c}} \log \frac{T}{K} \\ \min_{q\in(1/c,1]} \frac{T^q}{\min\{1,cq-1\}} \end{cases}\right) = \mathcal{O}\left(KT^{\frac{1}{c}} \log \frac{T}{K}\right).$$

B.1.2. Restless Setting

Let us now move to the restless setting. From Theorem 4.6, we have:

$$R_{\boldsymbol{\mu}} \leq 2K + T^{\frac{q}{q+1}} K \Upsilon_{\boldsymbol{\mu}} \left(\left\lceil \frac{T}{K} \right\rceil, q\right)^{\frac{1}{1+q}} \leq 2K + KT^{\frac{q}{q+1}} + K \begin{cases} \frac{T^{\frac{1-cq+q}{q+1}}}{K^{\frac{1-qc}{q+1}}(1-cq)} & \text{if } cq < 1 \\ T^{\frac{q}{q+1}} \left(\log \frac{T}{K}\right)^{\frac{1}{q+1}} & \text{if } cq = 1 \\ \frac{T^{\frac{q}{q+1}}}{cq-1} & \text{if } cq > 1 \end{cases}$$

$$\leq \mathcal{O}\left(K \begin{cases} \frac{T^{\frac{1-cq+q}{q+1}}}{K^{\frac{1-qc}{q+1}}(1-cq)} & \text{if } cq < 1 \\ T^{\frac{q}{q+1}} \left(\log \frac{T}{K}\right)^{\frac{1}{q+1}} & \text{if } cq = 1 \\ \frac{T^{\frac{q}{q+1}}}{\min\{1,cq-1\}} & \text{if } cq > 1 \end{cases}\right), \qquad \forall q \in [0,1].$$

For the case $c \in (0, 1)$, we consider the first case only and minimize over q:

$$R_{\boldsymbol{\mu}} \leq \mathcal{O}\left(K \min_{q\in[0,1]} \frac{T^{\frac{1-cq+q}{q+1}}}{K^{\frac{1-qc}{q+1}}(1-cq)}\right) \leq \mathcal{O}\left(\frac{K^{\frac{1+c}{2}} T^{1-\frac{c}{2}}}{1-c}\right),$$

for sufficiently large $T \gg K$. For the case $c = 1$, it is simple to prove that the case $cq = 1$ leads to the smallest regret:

$$R_{\boldsymbol{\mu}} \leq KT^{\frac{1}{c+1}} \left(\log \frac{T}{K}\right)^{\frac{c}{c+1}}.$$

Finally, for the case $c \in (1, +\infty)$, we have to consider all the three cases:

$$R_{\boldsymbol{\mu}} \leq \mathcal{O}\left(K \begin{cases} \min_{q\in[0,1/c)} \frac{T^{\frac{1-cq+q}{q+1}}}{K^{\frac{1-qc}{q+1}}(1-cq)} \\ T^{\frac{1}{c+1}} \left(\log \frac{T}{K}\right)^{\frac{c}{c+1}} \\ \min_{q\in(1/c,1]} \frac{T^{\frac{q}{q+1}}}{\min\{1,cq-1\}} \end{cases}\right) = KT^{\frac{1}{c+1}} \left(\log \frac{T}{K}\right)^{\frac{c}{c+1}}.$$

B.2. Efficient Update

Under the assumption that the window size depends on the number of pulls only and that $0 \leq h(n+1) - h(n) \leq 1$, we can employ the following efficient $\mathcal{O}(1)$ update for `R-ed-UCB`and `R-less-UCB`. Denoting with n the number of pulls of arm i, we update the estimator at every time step $t \in [T]$ as:

$$\widehat{\mu}_i^{h(n)}(t) = \frac{1}{h(n)} \left(a_n + \frac{t(a_n - b_n)}{h(n)} - \frac{c_n - d_n}{h(n)} \right),$$

where the following sequences are updated only when the arm is pulled:

$$a_n = \begin{cases} a_{n-1} + r_i(n) - r_i(n - h(n)) & \text{if } h(n) = h(n-1) \\ a_{n-1} + r_i(n) & \text{otherwise} \end{cases},$$

$$b_n = \begin{cases} b_{n-1} + r_i(n - h(n)) - r_i(n - 2h(n)) & \text{if } h(n) = h(n-1) \\ b_{n-1} + r_i(n - 2h(n) + 1) & \text{otherwise} \end{cases},$$

$$c_n = \begin{cases} c_{n-1} + nr_i(n) - (n - h(n))r_i(n - h(n)) & \text{if } h(n) = h(n-1) \\ c_{n-1} + nr_i(n) & \text{otherwise} \end{cases},$$

$$d_n = \begin{cases} d_{n-1} + (n - h(n))r_i(n - h(n)) - (n - 2h(n))r_i(n - 2h(n)) & \text{if } h(n) = h(n-1) \\ d_{n-1} + (n - 2h(n) + 1)r_i(n - 2h(n) + 1) & \text{otherwise} \end{cases},$$

where we have abbreviated $r_i(n) \coloneqq R_{t_{i,n}}$.

C | Algorithms tuning

The choices of the parameters of the baseline algorithms we compared `R-less/ed-UCB` with are the following:

- `Rexp3`: $V_T = K$ since in our experiments we consider the reward of each arm to evolve from 0 to 1, thus the maximum global variation possible is equal the number of arms of the bandit; $\gamma = \min\left\{1, \sqrt{\frac{K \log K}{(e-1)\Delta_T}}\right\}$; $\Delta_T = \lceil (K \log K)^{1/3} (T/V_T)^{2/3} \rceil$ as recommended by Besbes et al. [11];
- `KL-UCB`: $c = 3$ as required by the theoretical results on the regret provided by Garivier and Cappé [21];
- `Ser4`: according to what suggested by Allesiardo et al. [5] we selected $\delta = 1/T$, $\epsilon = \frac{1}{KT}$, and $\phi = \sqrt{\frac{N}{TK \log(KT)}}$;
- `SW-UCB`: as suggested by Garivier and Moulines [23] we selected the sliding-window $\tau = 4\sqrt{T \log T}$ and the constant $\xi = 0.6$;
- `SW-KL-UCB`as suggested by Garivier and Moulines [23] we selected the sliding-window $\tau = \sigma^{-4/5}$;
- `SW-TS`: as suggested by Trovò et al. [56] for the smoothly changing environment we set $\beta = 1/2$ and $\tau = T^{1-\beta} = \sqrt{T}$.

List of Figures

List of Tables

List of Algorithms

List of Symbols

Symbol	Description
log	Natural Logarithm
$\mathcal{B}(\cdot)$	Bernoulli Distribution
$\mathcal{N}(\cdot, \cdot)$	Gaussian Distribution
$\mathcal{U}(\cdot, \cdot)$	Uniform Distribution
$\mathcal{O}(\cdot)$	Big O,
$\mathbb{E}[\cdot]$	Expected Value
$\mathcal{O}(\cdot)$	Big O,
$\widetilde{\mathcal{O}}(\cdot)$	$\mathcal{O}$, disregarding logarithmic terms
$\boldsymbol{\nu}$	Instance of Multi-Armed-Bandit
T	Horizon of the bandit problem
K	Arms of the bandit instance
$\boldsymbol{\mu}$	Payoffs of the arms of the bandit
$\gamma_i(\cdot)$	Growth of arm i payoff
$\mathcal{H}$	History of the pulls and rewards
π	Policy
$\pi^*_{\boldsymbol{\mu},T}$	Optimal Policy for bandit $\boldsymbol{\mu}$ on horizon T
$N_{i,t}$	Number of pulls of arm i up to round t
$J_{\boldsymbol{\mu}}$	Cumulative Expected Reward
$R_{\boldsymbol{\mu}}(\pi, T)$	Regret suffered by policy π on bandit $\boldsymbol{\mu}$ on horizon T