Stochastic Rising Bandits

Alberto Maria Metelli [1] **Francesco Trovò** [1] **Matteo Pirola** [1] **Marcello Restelli** [1]

Abstract

This paper is in the field of stochastic Multi-Armed Bandits (MABs), i.e., those sequential selection techniques able to learn online using only the feedback given by the chosen option (a.k.a. arm). We study a particular case of the rested and restless bandits in which the arms' expected payoff is monotonically non-decreasing. This characteristic allows designing specifically crafted algorithms that exploit the regularity of the payoffs to provide tight regret bounds. We design an algorithm for the rested case (`R-ed-UCB`) and one for the restless case (`R-less-UCB`), providing a regret bound depending on the properties of the instance and, under certain circumstances, of $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$. We empirically compare our algorithms with state-of-the-art methods for non-stationary MABs over several synthetically generated tasks and an online model selection problem for a real-world dataset. Finally, using synthetic and real-world data, we illustrate the effectiveness of the proposed approaches compared with state-of-the-art algorithms for the non-stationary bandits.

1. Introduction

The classical stochastic MAB framework (Lattimore & Szepesvári, 2020) has been successfully applied to a number of applications, such as advertising, recommendation, and networking. MABs model the scenario in which a learner sequentially selects (a.k.a. pulls) an option (a.k.a. arm) in a finite set, and receives a feedback (a.k.a. reward) corresponding to the chosen option. The goal of online learning algorithms is to guarantee the *no-regret* property, meaning that the loss due to not knowing the best arm is increasing sublinearly with the number of pulls. One of the assumptions that allows designing no-regret algorithms consists in requiring that the payoff (a.k.a. expected reward) provided by the available options is *stationary*, i.e., rewards come from a fixed distribution.

However, the arms' payoff may change over time due to intrinsic modifications of the arms or the environment. A no-regret approach is offered by the *adversarial* algorithms, in which no assumption on the nature of the reward is required. It has been shown that, in this setting, it is possible to design effective algorithms, e.g., `EXP3` (Auer et al., 1995). However, in practice, their performance is unsatisfactory because the *non-stationarity* of real-world cases is far from being adversarial. Instead, non-stationarity is explicitly accounted for by a surge of methods that consider either abrupt changes (e.g., Garivier & Moulines, 2011), smoothly changing environments (e.g., Trovò et al., 2020) or bounded reward variation (e.g., Besbes et al., 2014).

While in non-stationary MABs the arms' payoff changes *naturally* over time, a different setting arises when the payoff changes as an effect of *pulling* the arm. This is the case of *rotting* bandits (Levine et al., 2017; Seznec et al., 2019), in which the payoff of the arms are monotonically non-increasing over the pulls, modeling degradation phenomena. Knowing the monotonicity property allows deriving more specialized algorithms, exploiting the process characteristics and further decreasing the regret w.r.t. unrestricted cases. Notably, the symmetric problem of monotonically non-decreasing payoffs cannot be addressed with the same approaches. Indeed, it was shown that it represents a significantly more complex problem, even for deterministic arms (Heidari et al., 2016). In this non-decreasing setting, a common assumption is the concavity of the payoff function that defines the *rising* bandits setting (Li et al., 2020).

The goal of this paper is to study the *stochastic* MAB problem when the arms' payoff is monotonically non-decreasing. This setting arises in several real-world sequential selection problems. For instance, suppose we have to choose among a set of optimization algorithms to maximize an unknown stochastic concave function. In this setting, we expect that all the algorithms progressively *increase* (on average) the function value and eventually converge to an optimal value, possibly with different speeds. Therefore, we wonder which candidate algorithm to assign the available resources (e.g., computational power or samples) to identify the one that converges faster to the optimum. This *online model selec-*

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy. Correspondence to: Alberto Maria Metelli <albertomaria.metelli@polimi.it>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

tion process can be modeled as a *rested* MAB (Tekin & Liu, 2012), like the rotting bandits (Levine et al., 2017), but with non-decreasing payoffs. Indeed, each optimization algorithm (arms) and the function value does not evolve if we do not select (pull) it. Another example that shows a non-decreasing expected reward is the selection of athletes for competitions. Athletes train in parallel and increase (on average) their performance. However, if participation in competitions is allowed to one athlete only, the trainer should select the one who has achieved the best performance so far. This problem is akin to the *restless* case (Tekin & Liu, 2012), like non-stationary bandits (Besbes et al., 2014), but with the additional assumption that payoffs are non-decreasing. Indeed, the athletes (arms) are evolving even if they are not selected (pulled) to compete.

Original Contribution In this paper, we study the stochastic rising bandits, i.e., stochastic bandits in which the payoffs are monotonically non-decreasing and concave, in both restless and rested formulations. More specifically:

- we show that the rested bandit with non-decreasing payoffs is *non-learnable*, i.e., the loss due to learning is linear with the number of pulls, unless additional assumptions on the payoff functions are enforced (e.g., concavity);
- we design `R-ed-UCB` and `R-less-UCB`, optimistic algorithms for the rising rested and restless bandits;
- we show that `R-ed-UCB` and `R-less-UCB` suffer an expected regret that depends on the payoff function profile and, under some conditions, of order $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$;[1]
- we illustrate, using synthetic and real-world data, the effectiveness of our approaches, compared with state-of-the-art algorithms for the non-stationary (restless) bandits.

2. Related Works

Restless and Rested Bandits The *rested* and *restless* bandit settings have been introduced by Tekin & Liu (2012) and further developed by (Ortner et al., 2012; Russac et al., 2019) in the restless version and by (Mintz et al., 2020; Pike-Burke & Grunewalder, 2019) in the rested one. Originally the evolution of the payoff was modeled via a suitable process, e.g., a Markov chain with finite state space or a linear regression process. For instance, Wang et al. (2020) proposes an optimistic approach based on the estimation of the transition kernel of the underlying chain. More recently, the terms rested and restless have been employed to denote arms whose payoff changes as time passes, for restless ones, or whenever being pulled, for rested ones (Seznec et al., 2019; 2020). That is the setting we target in this work.

Non-Stationary Bandits The restless bandits, without a fixed temporal reward evolution, are usually addressed via non-stationary MAB approaches, that include both passive (e.g., Garivier & Moulines, 2011; Besbes et al., 2014; Auer et al., 2019; Trovò et al., 2020) and active (e.g., Liu et al., 2018; Besson et al., 2019; Cao et al., 2019) methods. 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 and use only data gathered after the last change. Garivier & Moulines (2011) employ a discounted reward approach (`D-UCB`) or an adaptive sliding window (`SW-UCB`), proving a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret when the number of abrupt changes is known. Similar results have been obtained by (Auer et al., 2019) without knowing the number of changes, at the price of resorting to the doubling trick. (Besbes et al., 2014) provides an algorithm, namely `REXP3`, a modification `EXP3`, originally designed for adversarial MABs, to give a regret bound of $\mathcal{O}(T^{\frac{2}{3}})$ under the assumption that the total variation V_T of the arms' expected reward is known. The knowledge of V_T has been removed by Chen et al. (2019) using the doubling trick. In Trovò et al. (2020), an approach in which the combined use of a sliding window on a Thompson Sampling-like algorithm provides theoretical guarantees both on abruptly and smoothly changing environments. Nonetheless, in our setting, their result might lead to linear regret for specific instances. Notably, none of the above explicitly use assumptions on the monotonicity of the payoff over time.

Rising Bandits The *rising* bandit problem has been tackled in its deterministic version by (Heidari et al., 2016; Li et al., 2020). In Heidari et al. (2016), the authors design an online algorithm to minimize the regret of selecting an increasing and concave function among a finite set. This study assumes that the learner receives feedback about the true value of the reward function, i.e., no stochasticity is present. In Li et al. (2020), the authors model the problem of parameter optimization for machine learning models as a rising bandit setting. They propose an online algorithm having good empirical performance, still in the case of deterministic rewards. A case where the reward is increasing in expectation (or equivalently decreasing in loss), but no longer deterministic, is provided by Cella et al. (2021). However, the payoff follows a given parametric form known to the learner, who estimates such parameters in the best-arm identification and regret-minimization frameworks. The need for knowing the parametric form of the payoff makes these approaches hardly applicable for arbitrary increasing functions.

Corralling Bandits It is also worth mentioning the *corralling* bandits (Agarwal et al., 2017; Pacchiano et al., 2020b; Abbasi-Yadkori et al., 2020; Pacchiano et al., 2020a; Arora et al., 2021), a setting in which the goal is to minimize the regret of a process choosing among a finite set of bandit algorithms. This setting, close to online model selection, is characterized by particular assumptions. Indeed, each arm corresponds to a learning algorithm, operating on a bandit, endowed with a (possibly known) regret bound, sometimes

[1] With $\widetilde{\mathcal{O}}(\cdot)$ we disregard logarithmic terms in the order.

requiring additional conditions (e.g., stability).

3. Problem Setting

A K-armed MAB (Lattimore & Szepesvári, 2020) is defined as a vector of probability distributions $\boldsymbol{\nu}=(\nu_i)_{i\in[K]}$, where $\nu_i:\mathbb{N}^2\to\Delta(\mathbb{R})$ depends on a pair of parameters $(t,n)\in\mathbb{N}^2$ for every $i\in[K]$, where $[K]\coloneqq\{1,\dots,K\}$. Let $T\in\mathbb{N}$ be the optimization horizon, at each round $t\in[T]$, the agent selects an arm $I_t\in[K]$ and observes a reward $R_t\sim\nu_{I_t}(t,N_{I_t,t})$, where $N_{i,t}=\sum_{l=1}^{t}\mathbb{1}\{I_l=i\}$ is the number of times arm $i\in[K]$ was pulled up to round t. Thus, the reward depends, in general, on the current round t and on the number of pulls $N_{I_t,t}=N_{I_t,t-1}+1$ of arm I_t up to t. For every arm $i\in[K]$, we define its payoff $\mu_i:\mathbb{N}^2\to\mathbb{R}$ as the expectation of the reward, i.e., $\mu_i(t,n)=\mathbb{E}_{R\sim\nu_i(t,n)}[R]$ and denote the vector of payoffs as $\boldsymbol{\mu}=(\mu_i)_{i\in[K]}$. We assume that the payoffs are bounded in $[0,1]$, and that the rewards are σ^2-subgaussian, i.e., $\mathbb{E}_{R\sim\nu_i(t,n)}[e^{\lambda(R-\mu_i(t,n))}]\leqslant e^{\frac{\sigma\lambda^2}{2}}$, for every $\lambda\in\mathbb{R}$.

Rested and Restless Arms We revise the definition of *rested* and *restless* arms (Tekin & Liu, 2012).[2]

Definition 3.1 (Rested and Restless Arms)**.** *Let $\boldsymbol{\nu}$ be a MAB and let $i\in[K]$ be an arm, we say that:*

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

A K-armed bandit is rested (resp. restless) if all of its arms are rested (resp. restless).

Thus, the payoff of a rested arm changes when being pulled and, therefore, it models phenomena that evolve as a consequence of the agent intervention. Instead, a restless arm is in all regards a non-stationary arm (Besbes et al., 2014), and it is suitable for modeling a natural phenomenon that evolves for time passing, independently of the agent intervention.

Rising Bandits We revise the *rising* bandits notion, i.e., MABs with payoffs *non-decreasing* and *concave* as a function of (t,n) (Heidari et al., 2016).[3]

Assumption 3.1 (Non-Decreasing Payoff)**.** *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*:*

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

From an economic perspective, $\gamma_i(\cdot)$ represents the *increase of total return* (or payoff) we obtain by adding a factor of production, i.e., pulling the arm (rested) or letting time evolve for a unit (restless). In the next sections, we analyze how the following assumption defines a remarkable class of bandits with non-decreasing payoffs (Heidari et al., 2016).

Assumption 3.2 (Concave Payoff)**.** *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.:*

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

As pointed out by Heidari et al. (2016), the concavity assumption corresponds, in economics, to the *decrease of marginal returns* that emerges when adding a factor of production, i.e., pulling the arm (rested) or letting time evolve for one unit (restless).

Formally, we define *rising* a stochastic MAB in which both Assumption 3.1 and Assumption 3.2 hold.

Learning Problem Let $t\in[T]$ be a round, we denote with $\mathcal{H}_t=(I_l,R_l)_{l=1}^{t}$ the *history* of observations up to t. A (non-stationary) deterministic policy is a function $\pi:\mathcal{H}_{t-1}\mapsto I_t$ mapping a history to an arm, that is abbreviated as $\pi(t)\coloneqq\pi(\mathcal{H}_{t-1})$. The performance of a policy π in a MAB with payoffs $\boldsymbol{\mu}$ is the *expected cumulative reward* collected over the T rounds, formally:

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

and the expectation is computed over the histories. A policy $\pi^*_{\boldsymbol{\mu},T}$ is *optimal* if it maximizes the expected cumulative reward: $\pi^*_{\boldsymbol{\mu},T}\in\operatorname{argmax}_\pi\{J_{\boldsymbol{\mu}}(\pi,T)\}$. Denoting with $J^*_{\boldsymbol{\mu}}(T)\coloneqq J_{\boldsymbol{\mu}}(\pi^*_{\boldsymbol{\mu},T},T)$ the expected cumulative reward of an optimal policy, the suboptimal policies π are evaluated via the *expected cumulative regret*:

$$R_{\boldsymbol{\mu}}(\pi,T)\coloneqq J^*_{\boldsymbol{\mu}}(T)-J_{\boldsymbol{\mu}}(\pi,T). \tag{1}$$

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

$$\Upsilon_{\boldsymbol{\mu}}(M,q)\coloneqq\max_{i\in[K]}\left\{\sum_{l=1}^{M-1}\gamma_i(l)^q\right\}. \tag{2}$$

The cumulative increment accounts for how fast the payoffs reach their asymptotic value, i.e., become stationary. Intuitively, small values of $\Upsilon_{\boldsymbol{\mu}}(M,q)$ lead to simpler problems, as they are closer to stationary bandits. Table 1 reports

[2] We refer to the definition of (Levine et al., 2017; Seznec et al., 2020) and not to the one of (Tekin & Liu, 2012) that assumes an underlying Markov chain governing the arms' distributions.

[3] Deterministic bandits with non-decreasing payoffs were introduced in (Heidari et al., 2016) with the term *improving*. In (Li et al., 2020), the term *rising* was used to denote the improving bandits with concave payoffs (concavity was already employed by Heidari et al. (2016)).

Table 1. $\mathcal{O}$ rates of $\Upsilon_{\boldsymbol{\mu}}(M,q)$ in the case $\gamma_i(l)\leqslant f(l)$ for all $i\in[K]$ and $l\in\mathbb{N}$ (see also Lemma C.6).

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

some bounds on $\Upsilon_{\boldsymbol{\mu}}(M,q)$ for particular choices of $\gamma_i(l)$ and q. When $q=1$, the cumulative increment resembles the bounded variation $V_T \coloneqq \sum_{l=1}^{T-1}\max_{i\in[K]}\{\gamma_i(l)\}$ (Besbes et al., 2014), but $\Upsilon_{\boldsymbol{\mu}}(T,1)$ is smaller than V_T as the maximization over the arms appears outside the summation.

In the next sections, we devise and analyze learning algorithms for rested (Section 4) and restless (Section 5) rising bandits. We will present *optimistic* algorithms, whose structure is summarized in Algorithm 1 and parametrized by an exploration index $B_i(t)$ that will be designed case by case.

4. Stochastic Rising Rested Bandits

In this section, we consider the *Rising rested* bandits (R-ed) setting in which the arms' expected payoff increases only when it is pulled, i.e., $\mu_i(t,N_{i,t})\equiv\mu_i(N_{i,t})$.[4]

Oracle Policy We recall that the *oracle constant* policy, that always plays at each round $t\in[T]$ the arm that maximizes the sum of the payoffs over the horizon T, is optimal for the non-increasing rested bandits.

Theorem 4.1 (Heidari et al., 2016)**.** *Let $\pi^c_{\boldsymbol{\mu},T}$ be the* oracle constant *policy:*

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

Then, $\pi^c_{\boldsymbol{\mu},T}$ is optimal for the rested non-decreasing bandits (i.e., under Assumption 3.1).

The result holds under the non-decreasing property (Assumption 3.1) only, without requiring concavity (Assumption 3.2). However, this policy cannot be used in practice as it requires knowing the full function $\mu_i(\cdot)$ in advance.

4.1. Non-Learnability

We now prove a result highlighting the "hardness" of the non-decreasing rested bandits. We show that with no assumptions on the payoff $\mu_i(n)$ (e.g., concavity), it is impossible to devise a no-regret algorithm.

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

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

The intuition behind this result is that, if we enforce no condition on the increment $\gamma_i(n)$ we cannot predict how much the arm payoff will increase in the future. Therefore, we face the dilemma of whether or not to pull an arm that is currently believed to be suboptimal, hoping its payoff will increase. If we decide to pull it and its payoff will not actually increase, or if we decide not to pull it and its payoff will actually increase, becoming optimal, we will suffer linear regret. Thus, Theorem 4.2 highlights the importance of the concavity assumption (Assumption 3.2), providing an answer to an open question posed in (Heidari et al., 2016).

Remark 4.1 (About the Concavity Assumption)**.** *While without additional structure, e.g., concavity, the* non-decreasing rested *bandits are non-learnable (Theorem 4.2), the assumption is not necessary in other related settings. In particular,* non-decreasing restless *bandits are in all regard non-stationary bandits, for which no-regret algorithms exist under different assumptions about the number of change points (Garivier & Moulines, 2011) or a bounded total variation (Besbes et al., 2014). Furthermore, for* non-increasing rested *(rotting) bandits (Levine et al., 2017), a bounded payoff decrement between consecutive pulls is sufficient to devise a no-regret algorithm.*

Algorithm 1 R-□-UCB (□∈{less,ed})

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

4.2. Deterministic Setting

To progressively introduce the core ideas, we begin with the case of deterministic arms ($\sigma=0$). We devise an optimistic estimator of $\mu_i(t)$, namely $\overline{\mu}_i^{\text{R-ed}}(t)$, having observed the exact payoffs $(\mu_i(n))_{n=1}^{N_{i,t-1}}$. Differently from the rotting setting, these payoffs are an underestimation of $\mu_i(t)$. Therefore, we exploit the non-decreasing assumption (Assumption 3.1) to derive the identity:

$$\mu_i(t)=\underbrace{\mu_i(N_{i,t-1})}_{\text{(most recent payoff)}}+\underbrace{\sum_{n=N_{i,t-1}}^{t-1}\gamma_i(n)}_{\text{(sum of future increments)}}. \tag{3}$$

[4] We are employing the original definition of rested arms of (Levine et al., 2017) in which $\mu_i(n)$ is the payoff of arm i when it is pulled *for the n-th time*.

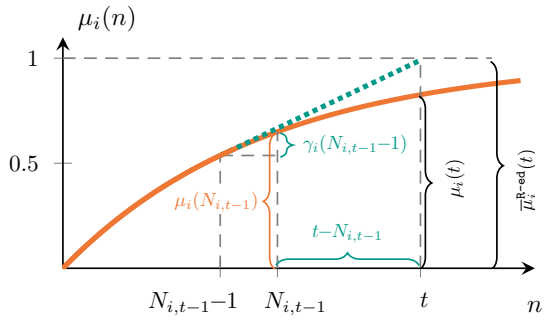


Figure 1. Graphical representation of the estimator construction $\overline{\mu}_i^{\text{R-ed}}(t)$ for the rested deterministic setting.

By exploiting the concavity (Assumption 3.2), 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, leading to the following estimator:

$$\overline{\mu}_i^{\text{R-ed}}(t) := \underbrace{\mu_i(N_{i,t-1})}_{\text{(most recent payoff)}} + (t-N_{i,t-1}) \underbrace{\gamma_i(N_{i,t-1}-1)}_{\text{(most recent increment)}}, \quad (4)$$

if $N_{i,t-1} \geqslant 2$ else $\overline{\mu}_i^{\text{R-ed}}(t) := +\infty$. Figure 1 illustrates the construction of the estimator. The optimism of $\overline{\mu}_i^{\text{R-ed}}$ and a bias bound are proved in Lemma A.2.

Regret Analysis We are now ready to provide the regret analysis of R-ed-UCB, i.e., Algorithm 1 when we employ as exploration index $B_i(t) \equiv \overline{\mu}_i^{\text{R-ed}}(t)$.

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

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

The regret depends on a parameter $q \in [0,1]$ that can be selected to tighten the bound, whose optimal value depends on $\Upsilon_{\boldsymbol{\mu}}(\cdot, q)$, that is a function on the horizon T. Some examples, when $\gamma_i(t) \leqslant l^{-c}$ for $c>0$, are reported in Figure 2.

4.3. Stochastic Setting

Moving to the R-ed stochastic setting ($\sigma > 0$), we cannot directly exploit the estimator in Equation (4). Indeed, we only observe the sequence of noisy rewards $(R_{t_{i,n}})_{n=1}^{N_{i,t-1}}$, where $t_{i,n} \in [T]$ is the round at which arm $i \in [K]$ was pulled for the n-th time. To cope with stochasticity, we need to employ an h-wide window made of the h most recent samples, similarly to what has been proposed by Seznec et al. (2020). The choice of h represents a *bias-variance trade-off* between employing few recent observations (less biased), compared to many past observations (less variance). For

	$c \leqslant 1$	$c \geqslant 1$
Rested	T	$KT^{\frac{1}{c}}$
Restless	$K^{\frac{1+c}{2}} T^{1-\frac{c}{2}}$	$KT^{\frac{1}{c+1}}$

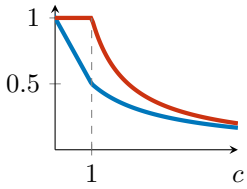


Figure 2. Regret bounds $\widetilde{\mathcal{O}}$ rates optimized over q for R-less and R-ed deterministic bandits when $\gamma_i(l) \leqslant l^{-c}$ for $c>0$.

$h \in [N_{i,t-1}]$, the resulting estimator $\widehat{\mu}_i^{\text{R-ed},h}(t)$ is given by:

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

if $h \leqslant \lfloor N_{i,t-1}/2 \rfloor$, else $\widehat{\mu}_i^{\text{R-ed},h}(t) := +\infty$. The construction of the estimator is shown in Appendix A.1 and relies on the idea of averaging several estimators of the form of Equation (4) instanced using as starting points different number of pulls $N_{i,t-1}-l+1$ for $l \in [h]$ and replacing the true payoff with the corresponding reward sample. An efficient way to compute this estimator is reported in Appendix D.

Regret Analysis By making use of the presented estimator, we build the following optimistic exploration index:

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

$$\beta_i^{\text{R-ed},h_{i,t}}(t,\delta_t) := \sigma(t - N_{i,t-1} + h_{i,t} - 1) \sqrt{\frac{10 \log \frac{1}{\delta_t}}{h_{i,t}^3}},$$

and $h_{i,t}$ are arm-and-time-dependent window sizes and δ_t is a time-dependent confidence parameter. By choosing the window size depending linearly on the number of pulls, we are able to provide the following regret bound.

Theorem 4.4. *Let $T \in \mathbb{N}$, then R-ed-UCB (Algorithm 1) with $B_i(t) \equiv \widehat{\mu}_i^{\text{R-ed},h_{i,t}}(t) + \beta_i^{\text{R-ed},h_{i,t}}(t)$, $h_{i,t} = \lfloor \epsilon N_{i,t-1} \rfloor$ 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:*

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

This result deserves some comments. First, compared with the corresponding deterministic R-ed regret bound (Theorem 4.3), it reflects a similar dependence of the cumulative

increment $\Upsilon_{\boldsymbol{\mu}}$, although it now involves the ϵ parameter defining the window size $h_{i,t}=\lfloor \epsilon N_{i,t-1} \rfloor$. Second, it includes an additional term of order $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$ that is due to the noise σ presence that increases inversely w.r.t. the ϵ.[5] Thus, we visualize a trade-off in the choice of ϵ: larger windows ($\epsilon \approx 1$) are beneficial for the first term, but they enlarge the constant $1/(1-2\epsilon)$ multiplying the second component.

Remark 4.2 (Comparison with Adversarial Bandits)**.** *The* `R-ed` *setting can be mapped to an* adversarial *bandit (Auer et al., 2002) with an* adaptive *(i.e., non-oblivious) adversary. Indeed, the arm payoff $\mu_i(N_{i,t})$ can be thought to as selected by an adversary who has access to the previous learner choices (i.e., the history $\mathcal{H}_{t-1}$), specifically to the number of pulls $N_{i,t}$. However, although adversarial bandit algorithms, such as* `EXP3` *(Auer et al., 2002) and* `OSMD` *(Audibert et al., 2014), suffer $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, these results are not comparable with ours. Indeed, while these correspond to guarantees on the* external regret*, the regret definition we employ in Section 3 is a notion of* policy regret *(Dekel et al., 2012).*

5. Stochastic Rising Restless Bandits

In this section, we consider the *Rising restless* bandits (`R-less`) in which the payoff increases at every round regardless the arm is pulled, i.e., $\mu_i(t,N_{i,t}) \equiv \mu_i(t)$.

Oracle Policy We start recalling that the *oracle greedy* policy, i.e., the policy selecting at each round $t \in [T]$ the arm with largest payoff, is optimal for the non-decreasing restless bandit setting.

Theorem 5.1 (Seznec et al., 2020)**.** *Let $\pi^g_{\boldsymbol{\mu}}$ be the* oracle greedy *policy:*

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

Then, $\pi^g_{\boldsymbol{\mu}}$ is optimal for the restless non-decreasing bandits (i.e., under Assumption 3.1).

Notice that $\pi^g_{\boldsymbol{\mu}}$ is optimal under the non-decreasing payoff assumption (Assumption 3.1) only, without requiring the concavity (Assumption 3.2). We can now first appreciate an important difference between *rising* and *rotting* bandits. While for the rotting bandits the oracle *greedy* policy is optimal for both the rested and restless settings, for the rising bandits it remains optimal in the restless case only. Indeed, for the rising rested case, as shown in Theorem 4.1, the oracle *constant* policy is needed to achieve optimality.

[5] In particular, when $\gamma_i(n)$ decreases sufficiently fast (see Table 1), the regret is dominated by the $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$ component.

5.1. Deterministic Setting

We begin with the case of deterministic arms ($\sigma = 0$). Similarly to the rested case, we design an optimistic estimator of $\mu_i(t)$, namely $\overline{\mu}_i^{\texttt{R-less}}(t)$, employing the exact payoffs $(\mu_i(t_{i,n}))_{n=1}^{N_{i,t-1}}$. To this end, we exploit the non-decreasing assumption (Assumption 3.1) 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 leverage the concavity (Assumption 3.2) 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, leading to the estimator:

$$\overline{\mu}_i^{\texttt{R-less}}(t) := \underbrace{\mu_i(t_{i,N_{i,t-1}})}_{\text{(most recent payoff)}} + (t - t_{i,N_{i,t-1}}) \underbrace{\frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}}}_{\text{(most recent increment)}}, \tag{5}$$

if $N_{i,t-1} \geqslant 2$, else $\overline{\mu}_i^{\texttt{R-less}}(t) := +\infty$. Lemma A.5 shows that $\overline{\mu}_i^{\texttt{R-less}}$ is optimistic and provides a bias bound.

Regret Analysis We now provide the regret analysis of `R-less-UCB` that is obtained from Algorithm 1, when setting $B_i(t) \equiv \overline{\mu}_i^{\texttt{R-less}}(t)$.

Theorem 5.2. *Let $T \in \mathbb{N}$, then* `R-less-UCB` *(Algorithm 1) with $B_i(t) \equiv \overline{\mu}_i^{\texttt{R-less}}(t)$ suffers an expected regret bounded, for every $q \in [0,1]$, as:*

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

Similarly to Theorem 4.3, the result depends on the free parameter $q \in [0,1]$, that can be chosen to tighten the bound. It is worth noting that the regret bound of the `R-less` deterministic case (Theorem 5.2) is always smaller than that of the `R-ed` deterministic case (Theorem 4.3). Indeed, ignoring the dependence on K, we have $R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) = \mathcal{O}\left(R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB}, T)^{\frac{1}{q+1}} \right)$. The following example clarifies the role of q for both the restless and rested case.

Example 5.1. *Suppose that for all $i \in [K]$, we have $\gamma_i(l) \leqslant l^{-c}$ for $c > 0$. The expressions of bounds on $\Upsilon_{\boldsymbol{\mu}}(\cdot, q)$ have been shown in Table 1. Different values of $q \in [0,1]$ should be selected to tighten the regret bounds depending on the value of c. Figure 2 reports the optimized bounds for the deterministic* `R-less` *and* `R-ed` *(derivation in Appendix B).*

5.2. Stochastic Setting

In the stochastic setting ($\sigma>0$), we have access to noisy versions of μ_i only, i.e., $(R_{t_{i,n}})_{n=1}^{N_{i,t-1}}$. Intuitively, we might be tempted to straightforwardly extend the derivation of the rested case by averaging h estimators like the ones in Equation (5), instanced with different time instants $t_{i,N_{i,t-1}}$. Unfortunately, this approach is unsuccessful for technical issues since the increment term would include the difference of time instants $t_{i,N_{i,t-1}}-t_{i,N_{i,t-1}-1}$ that, in the stochastic setting, are random variables correlated with the observed rewards $R_{t_{i,n}}$. For this reason, at the price of a larger bias, we employ the same estimator used in the stochastic rested case, defined for $h\in[N_{i,t-1}]$:

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

if $h_{i,t}\leqslant\lfloor N_{i,t-1}/2\rfloor$, else $\widehat{\mu}_i^{\texttt{R-less},h}(t)\coloneqq+\infty$. Additional details on the estimator construction is reported in Appendix A.2 together with its analysis.

Regret Analysis We provide the regret analysis of `R-less-UCB` when we employ the exploration index analogous to that of the rested case:

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

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

and $h_{i,t}$ are a arm-and-time-dependent window sizes and δ_t is a time-dependent confidence. The regret bound is given by the following result.

Theorem 5.3. *Let $T\in\mathbb{N}$, then `R-less-UCB` (Algorithm 1) 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:*

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

Some observations are in order. First, compared to the bound for the rested case in Theorem 4.4, we note the same dependence of $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$ due to the noise presence σ. Concerning the second term, compared with the one of the deterministic case (Theorem 5.2), we worsen the dependence on T and an inverse dependence on the ϵ and $1-2\epsilon$ parameters appear. This is due to the usage of the h-wide window instead of the last sample and that, all other things being equal, the estimator employed for the stochastic case, as already discussed, is looser compared to the one for the deterministic case. Finally, our result is not fully comparable with (Besbes et al., 2014) for generic non-stationary bandits with bounded variation because, as already mentioned, $\Upsilon_{\boldsymbol{\mu}}(\lceil T/K\rceil,q)$ may be smaller than V_T. Moreover, we achieve such a bound with no knowledge about $\Upsilon_{\boldsymbol{\mu}}$, while the work by (Besbes et al., 2014) requires knowing V_T.

6. Numerical Simulations

We numerically tested `R-less-UCB` and `R-ed-UCB` w.r.t. state-of-the-art algorithms for non-stationary MABs in the *restless* and *rested* settings, respectively.[6]

Algorithms We consider the following baseline algorithms: `Rexp3` (Besbes et al., 2014), a non-stationary MAB algorithm based on variation budget, `KL-UCB` (Garivier & Cappé, 2011), one of the most effective stationary MAB algorithms, `Ser4` (Allesiardo et al., 2017), which considers best arm switches during the process, and sliding-window algorithms such as `SW-UCB` (Garivier & Moulines, 2011), `SW-KL-UCB` (Combes & Proutiere, 2014), and `SW-TS` (Trovò et al., 2020) that are generally able to deal with non-stationary restless settings. The parameters for all the baseline algorithms have been set as recommended in the corresponding papers (see also Appendix E). For our algorithms, the window is set as $h_{i,t}=\lfloor\epsilon N_{i,t-1}\rfloor$ (as prescribed by Theorems 4.4 and 5.3). We remark that while the baseline algorithms are suited for the restless case, in the rested case, no algorithm has been designed to cope with the stochastic rising setting, provided that no knowledge on the payoff function is available. We compare the algorithms in terms of empirical cumulative regret $\widehat{R}_{\boldsymbol{\mu}}(\pi,t)$, which is the empirical counterpart of the expected cumulative regret $R_{\boldsymbol{\mu}}(\pi,t)$ at round t averaged over multiple independent runs.

6.1. Restless setting

To evaluate `R-less-UCB` in the restless setting, we run the aforementioned algorithms on a problem with $K=15$ arms over a time horizon of $T=200{,}000$ rounds, setting $\epsilon=1/4$. The payoff functions $\mu_i(\cdot)$ are chosen in these families: $F_{\text{exp}}=\{f(t)=c(1-e^{-at})\}$ and $F_{\text{poly}}=\{f(t)=c\big(1-b(t+b^{1/\rho})^{-\rho}\big)\}$, where $a,c,\rho\in(0,1]$ and $b\in\mathbb{R}_{\geqslant 0}$ are parameters, whose values have been selected randomly. By construction all functions $f\in F_{\text{exp}}\cup F_{\text{poly}}$ satisfy

[6] Details of the experimental setting, and additional results are provided in Appendix E. The code to reproduce the experiments is available at https://github.com/albertometelli/stochastic-rising-bandits.

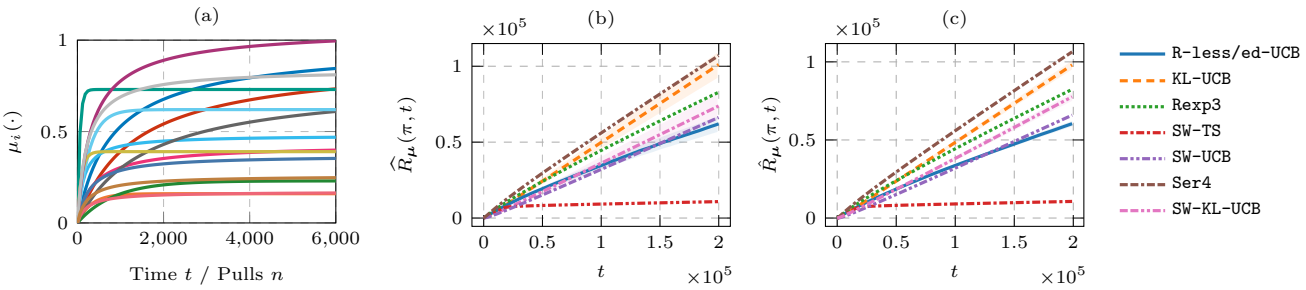


Figure 3. 15 arms bandit setting: (a) first 6000 rounds/pulls of the payoff functions, (b) cumulative regret in the `R-less` scenario, (c) cumulative regret in the `R-ed` scenario (100 runs 95% c.i.).

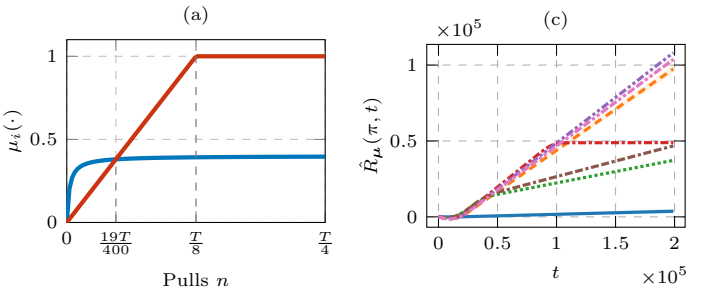


Figure 4. 2 arms `R-ed` bandit setting: (a) payoff functions, (b) cumulative regret (100 runs, 95% c.i.).

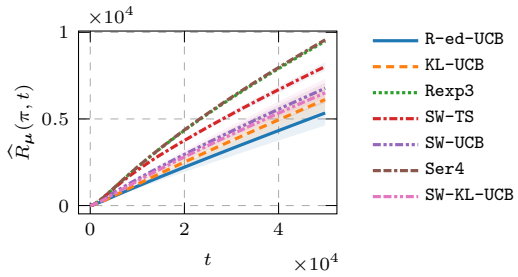


Figure 5. Cumulative regret in the online model selection on IMDB dataset (30 runs, 95% c.i.).

Assumptions 3.1 and 3.2. The functions coming from F_{exp} (exponential functions) have a sudden increase, while ones from F_{poly} (polynomial functions) have a slower growth rate, leading to different cumulative increments $\Upsilon_{\boldsymbol{\mu}}$. The stochasticity is realized by adding a Gaussian noise with $\sigma=0.1$. The generated functions are shown in Figure 3a.

The empirical cumulative regret $\widehat{R}_{\boldsymbol{\mu}}(\pi,t)$ is provided in Figure 3b. 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. As commonly happening in practice, TS-based approaches tend to outperform UCB ones. Indeed, `R-less-UCB` displays the second-best curve overall and achieves the best performance among the UCB-like algorithms.

6.2. Rested setting

We employ the same arms generated for the restless case to evaluate `R-ed-UCB` in the rested setting. We plot the empirical cumulative regret in Figure 3c. `SW-TS` is confirmed as the best algorithm at the end of the time horizon, although other algorithms (`SW-UCB` and `SW-KL-UCB`) suffer less regret at the beginning of learning. `R-ed-UCB` pays the price of the initial exploration, but at the end of the horizon, it manages to achieve the second-best performance. Notice that, besides `R-ed-UCB`, all other baseline algorithms are designed for the restless setting and are not endowed with any guarantee on the regret in the rested scenario.

To highlight this fact, we designed a particular 2-arms rising rested bandit in which the optimal arm reveals only when pulled a sufficient number of times (linear in T). The payoff functions, fulfilling Assumptions 3.1 and 3.2, are shown in Figure 4a and the algorithms' empirical regrets in Figure 4b. Note that in this setting the expected (instantaneous) regret may be negative for $t<\frac{19T}{400}$, and this is the case for most of the algorithms for $t<20{,}000$. While for the first $\approx 20{,}000$ rounds `R-ed-UCB` is on par with the other algorithms, it outperforms all the other policies over a longer run. Note that the regret for `Rexp3` and `Ser4` is decreasing the slope for $t>40{,}000$, meaning that they are somehow reacting to the change in the reward of the two arms. `SW-TS` starts reacting even later, at around $t\approx 100{,}000$. However, they are not prompt to detect such a change in the rewards and, therefore, collect a large regret in the first part of the learning process. The other algorithms suffer a linear regret at the end of the time horizon since they do not employ forgetting mechanisms or because the sliding window should be tuned knowing the characteristics of the expected reward.

6.3. IMDB dataset (rested)

We investigate the performance of R-ed-UCB on an *online model selection* task for a real-world dataset. We employ the IMDB dataset, made of $50{,}000$ reviews of movies (scores from 0 to 10). We preprocessed the data as done by Maas et al. (2011) to obtain a binary classification problem. Each review $\boldsymbol{x}_t$ lies in a $d=10{,}000$ dimensional feature space, where each feature is the frequency of the most common English words. Each arm corresponds to a different online optimization algorithm, i.e., two of them are Online Logistic Regression algorithms with different learning rate schemes, and the other five are Neural Networks with different topologies. We provide additional information on the arms of the bandit in Appendix E.2. At each round, a sample $\boldsymbol{x}_t$ is randomly selected from the dataset, a reward of 1 is generated for a correct classification, 0 otherwise, and, finally, the online update step is performed for the chosen algorithm.

The empirical regret is plotted in Figure 5. We can see that R-ed-UCB, with $\epsilon=1/32$ outperforms the considered baselines. Compared to the synthetic simulations, the smaller window choice is justified by the fact that we need to take into account that the average learning curves of the classification algorithms are not guaranteed to be non-decreasing nor concave on the single run. However, keeping the window linear in $N_{i,t-1}$ is crucial for the regret guarantees of Theorem 4.4.

7. Discussion and Conclusions

This paper studied the MAB problem when the payoffs are non-decreasing functions that evolve either when pulling the corresponding arm (rested) or for time passing (restless). We showed that, for the rested case, an assumption on the payoff (e.g., concavity) is essential to make the problem learnable. We presented novel algorithms that suitably employ the concavity assumption to build proper estimators for both settings. These algorithms are proven to suffer a regret made of a first instance-independent component of $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$ and an instance-dependent component involving the cumulative increment function $\Upsilon_{\boldsymbol{\mu}}(\cdot,q)$. For the rested setting, ours represent the first no-regret algorithm for the stochastic rising bandits. The experimental evaluation confirmed our theoretical findings showing advantages over state-of-the-art algorithms designed for non-stationary bandits, especially in the rested setting. The natural future research direction consists of studying the complexity of the learning problem in stochastic rising rested and restless bandits, focusing on deriving suitable regret lower bounds. Other future works include investigating the best-arm identification setting and, motivated by the online model selection, analysing the alternative case in which the optimization algorithms associated with the arms act on a shared vector of parameters.

References

Abbasi-Yadkori, Y., Pacchiano, A., and Phan, M. Regret balancing for bandit and RL model selection. *CoRR*, abs/2006.05491, 2020.

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

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

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

Audibert, J., Bubeck, S., and Lugosi, G. Regret in online combinatorial optimization. *Math. Oper. Res.*, 39(1):31–45, 2014.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Auer, P., Gajane, P., and Ortner, R. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Conference on Learning Theory (COLT)*, volume 99, pp. 138–158, 2019.

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

Besson, L., Kaufmann, E., Maillard, O.-A., and Seznec, J. Efficient change-point detection for tackling piecewise-stationary bandits. *arXiv preprint arXiv:1902.01575*, 2019.

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

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

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

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

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

Dekel, O., Tewari, A., and Arora, R. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Doob, J. L. *Stochastic processes*, volume 7. Wiley New York, 1953.

Garivier, A. and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 359–376, 2011.

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

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

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

Levine, N., Crammer, K., and Mannor, S. Rotting bandits. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Proceedings of the conference on Neural Information Processing Systems (NeurIPS)*, pp. 3074–3083, 2017.

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

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

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

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

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

Pacchiano, A., Dann, C., Gentile, C., and Bartlett, P. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020a.

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

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

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

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

Seznec, J., Ménard, P., Lazaric, A., and Valko, M. A single algorithm for both restless and rested rotting bandits. In *Proceedings of the international conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pp. 3784–3794, 2020.

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

Trovò, F., Paladino, S., Restelli, M., and Gatti, N. Sliding-window thompson sampling for non-stationary settings.

Journal of Artificial Intelligence Research, 68:311–364, 2020.

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

A. Proofs and Derivations

In this section, we provide the proof of the results presented in the main paper.

A.1. Proofs of Section 4

Theorem 4.1 (Heidari et al., 2016)**.** *Let $\pi^c_{\boldsymbol{\mu},T}$ be the* oracle constant *policy:*

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

Then, $\pi^c_{\boldsymbol{\mu},T}$ is optimal for the rested non-decreasing bandits (i.e., under Assumption 3.1).

Proof. The proof is reported in Proposition 1 of (Heidari et al., 2016). □

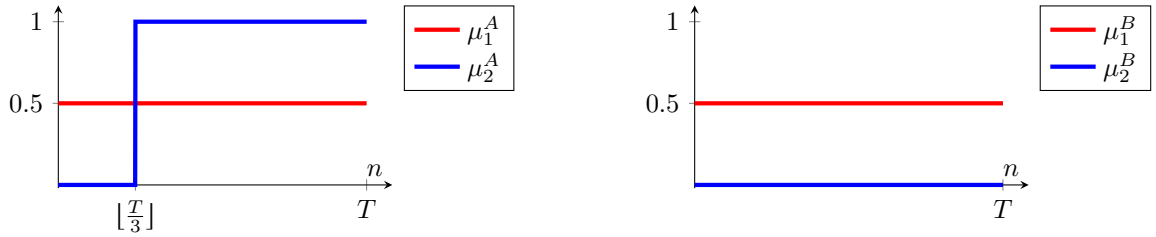
Lemma A.1. *In the noiseless ($\sigma=0$) setting, there exists a 2-armed non-increasing non-concave bandit such that any learning policy π suffers regret:*

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

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

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

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 $\lfloor \frac{T}{3} \rfloor$. Thus, we have:

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

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

$$J_{\boldsymbol{\mu}^A}(\pi,T) = \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\mu}^A}[N_{1,T}] + \max\left\{0, \mathop{\mathbb{E}}_{\boldsymbol{\mu}^A}[N_{2,T}] - \left\lfloor \frac{T}{3} \right\rfloor\right\} \tag{6}$$
$$= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\mu}^A}[N_{1,T}] + \max\left\{0, \left\lceil \frac{2}{3}T \right\rceil - \mathop{\mathbb{E}}_{\boldsymbol{\mu}^A}[N_{1,T}]\right\}, \tag{7}$$

where Equation (6) 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 (7) 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) : $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] \geqslant \lceil \frac{2}{3}T \rceil$

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

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

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

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

that is maximized by taking the minimum value of $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}]$ possible, that is $\mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,T}] \geqslant \mathbb{E}_{\boldsymbol{\mu}^A}[N_{1,\lfloor \frac{T}{3} \rfloor}] = n_1$. Putting all together, we have:

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

having observed that $n_1 \leqslant \lfloor \frac{T}{3} \rfloor$. 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:

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

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

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

having minimized over n_1. □

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

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

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

Lemma A.2. *For every arm $i \in [K]$ and every round $t \in [T]$, let us define:*

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

if $N_{i,t-1}\geqslant 2$ *else* $\overline{\mu}_i^{R\text{-}ed}(t):=+\infty$. *Then,* $\overline{\mu}_i^{R\text{-}ed}(t)\geqslant\mu_i(t)$ *and, if* $N_{i,t-1}\geqslant 2$, *it holds that:*

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

Proof. Let us consider the following derivation:

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

where the inequality holds thanks to Assumption 3.2, having observed that $\sum_{n=N_{i,t-1}}^{t-1}\gamma_i(n)\leqslant(t-N_{i,t-1})\gamma_i(N_{i,t-1})\leqslant(t-N_{i,t-1})\gamma_i(N_{i,t-1}-1)$. For the bias bound, when $N_{i,t-1}\geqslant 2$, we consider the following derivation:

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

having observed that $\mu_i(N_{i,t-1})\leqslant\mu_i(N_{i,t})$ by Assumption 3.1. □

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

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

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 $i^*\in\operatorname{argmax}_{i\in[K]}\left\{\sum_{l\in[T]}\mu_i(l)\right\}$. 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., $B_{i*}(t)\leqslant B_{I_t}(t)$:

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

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

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

where the inequality follows from Lemma A.2. We are going to decompose the summation of this term over the K arms:

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

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 $x\geqslant 0$ it holds that

$\min\{1,x\}\leqslant\min\{1,x\}^q\leqslant x^q$. By applying this latter inequality to the inner summation, we get:

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

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

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

where the last inequality is obtained from Lemma C.2. □

Estimator Construction for the Stochastic Rising Rested Setting Before moving to the proofs, we provide some intuition behind the estimator construction. We start observing that for every $l\in\{2,\dots,N_{i,t-1}\}$, we have that:

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

where the inequality follows from Assumption 3.2.[7] Since we do not have access to the exact payoffs $\mu_i(l)$ and exact increments $\gamma_i(l-1)=\mu_i(l)-\mu_i(l-1)$, one may be tempted to directly replace them with the corresponding point estimates $R_{t_{i,l}}$ and $R_{t_{i,l}}-R_{t_{i,l-1}}$ and average the resulting estimators for a window of the most recent h values of l. Unfortunately, while replacing $\mu_i(l)$ with $R_{t_{i,l}}$ is a viable option, replacing $\gamma_i(l-1)$ with $R_{t_{i,l}}-R_{t_{i,l-1}}$ will prevent concentration since the estimate $R_{i,t_l}-R_{i,t_{l-1}}$ is too unstable. To this end, before moving to the estimator, we need a further bounding step to get a more stable, although looser, quantity. Based on Lemma C.3, we bound for every $l\in\{2,\dots,N_{i,t-1}\}$ and $h\in[l-1]$:

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

We can now introduce the optimistic approximation of $\mu_i(t)$, i.e., $\widetilde{\mu}_i^{\texttt{R-ed},h}(t)$, and the corresponding estimator, i.e., $\widehat{\mu}_i^{\texttt{R-ed},h}(t)$, that are defined in terms of a window of size $1\leqslant h\leqslant\lfloor N_{i,t-1}/2\rfloor$:

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

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

The proof is composed of the following steps:

(i) Lemma A.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 w.r.t. $\mu_i(N_{i,t})$ for every value of h;

(ii) Lemma A.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 $h_{i,h}:=h(N_{i,t-1})$ is a function of the number of pulls $N_{i,t-1}$ only;

(iii) Theorem 4.4 bounds the expected regret of `R-ed-UCB` when $h_{i,h}=\lfloor\epsilon N_{i,t-1}\rfloor$, for $\epsilon\in(0,1/2)$.

[7] The estimator of the deterministic case in Equation (4) is obtained by setting $l=N_{i,t-1}$.

Lemma A.3. *For every arm $i\in[K]$, every round $t\in[T]$, and window width $1\leqslant h\leqslant\lfloor N_{i,t-1}/2\rfloor$, let us define:*

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

otherwise if $h=0$, we set $\widetilde{\mu}_i^{R\text{-}ed,h}(t):=+\infty$. Then, $\widetilde{\mu}_i^{R\text{-}ed,h}(t)\geqslant\mu_i(t)$ and, if $N_{i,t-1}\geqslant 2$, it holds that:

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

Proof. Following the derivation provided above, we have for every $l\in\{2,\dots,N_{i,t-1}\}$:

$$\begin{aligned}\mu_i(t)&=\mu_i(l)+\sum_{j=l}^{t-1}\gamma_i(j)\\&\leqslant\mu_i(l)+(t-l)\gamma_i(l-1)&(8)\\&\leqslant\mu_i(l)+(t-l)\frac{\mu_i(l)-\mu_i(l-h)}{h},&(9)\end{aligned}$$

where line (8) follows from Assumption 3.2, line (9) is obtained from Lemma C.3. By averaging over the most recent $1\leqslant h\leqslant\lfloor N_{i,t-1}/2\rfloor$ pulls, we obtain:

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

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

$$\begin{aligned}\widetilde{\mu}_i^{\text{R-ed},h}(t)-\mu_i(N_{i,t})&=\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(l)+(t-l)\frac{\mu_i(l)-\mu_i(l-h)}{h}\right)-\mu_i(N_{i,t})&(10)\\&\leqslant\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\frac{\mu_i(l)-\mu_i(l-h)}{h}\\&=\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\frac{1}{h}\sum_{j=l-h}^{l-1}\gamma_j(l)\\&\leqslant\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\gamma_i(l-h)&(11)\\&\leqslant\frac{1}{2}(2t-2N_{i,t-1}+h-1)\gamma_i(N_{i,t-1}-2h+1).&(12)\end{aligned}$$

where line (10) follows from Assumption 3.1 applied as $\mu_i(l)\leqslant\mu_i(N_{i,t})$, line (11) follows from Assumption 3.2 and bounding $\frac{1}{h}\sum_{j=l-h}^{l-1}\gamma_j(l)\leqslant\gamma_i(l-h)$ and line (12) is derived still from Assumption 3.2, $\gamma_i(l-h)\leqslant\gamma_i(N_{i,t-1}-2h+1)$ and computing the summation. □

Lemma A.4. *For every arm $i\in[K]$, every round $t\in[T]$, and window width $1\leqslant h\leqslant\lfloor N_{i,t-1}/2\rfloor$, let us define:*

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

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

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

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

Proof. First of all, we observe under the event $\{h_{i,t}=0\}$, then $\widehat{\mu}_i^{\text{R-ed},h_{i,t}}(t) = \widetilde{\mu}_i^{\text{R-ed},h_{i,t}}(t) = \beta_i^{\text{R-ed},h_{i,t}}(t,\delta_t) = +\infty$. By convening that $(+\infty)-(+\infty)=0$, we have that $0 > \beta_i^{\text{R-ed},h_{i,t}}(t,\delta_t)$ is not satisfied. Thus, we perform the analysis under the event $\{h_{i,t} \geqslant 1\}$. We first get rid of the dependence on the random number of pulls $N_{i,t-1}$:

$$\begin{aligned}
&\Pr\left(\left|\widehat{\mu}_i^{\text{R-ed},h_{i,t}}(t) - \widetilde{\mu}_i^{\text{R-ed},h_{i,t}}(t)\right| > \beta_i^{\text{R-ed},h_{i,t}}(t,\delta_t)\right) \\
&\qquad = \Pr\left(\left|\widehat{\mu}_i^{\text{R-ed},h(N_{i,t-1})}(t) - \widetilde{\mu}_i^{\text{R-ed},h(N_{i,t-1})}(t)\right| > \beta_i^{\text{R-ed},h(N_{i,t-1})}(t,\delta_t)\right) \qquad (13) \\
&\qquad \leqslant \Pr\left(\exists n \in \{0,\dots,t-1\} \text{ s.t. } h(n) \geqslant 1 : \left|\widehat{\mu}_i^{\text{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\text{R-ed},h(n)}(t)\right| > \beta_i^{\text{R-ed},h(n)}(t,\delta_t)\right) \\
&\qquad \leqslant \sum_{n \in \{0,\dots,t-1\} : h(n) \geqslant 1} \Pr\left(\left|\widehat{\mu}_i^{\text{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\text{R-ed},h(n)}(t)\right| > \beta_i^{\text{R-ed},h(n)}(t,\delta_t)\right), \qquad (14)
\end{aligned}$$

where line (13) derives from the definition of $h_{i,t} = h(N_{i,t-1})$ and line (14) 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:

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

where $X_l := R_{t_{i,l}} - \mu_i(l)$. 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 C.5 for weighted sums of subgaussian martingale difference sequences. To this purpose, we compute the sum of the square weights:

$$\begin{aligned}
&\sum_{l=n-h(n)+1}^{n} \left(1 + \frac{t-l}{h(n)}\right)^2 + \sum_{l=n-h(n)+1}^{n} \left(\frac{t-l}{h(n)}\right)^2 \\
&\qquad \leqslant h(n)\left(1 + \frac{t-n+h(n)-1}{h(n)}\right)^2 + h(n)\left(\frac{t-n+h(n)-1}{h(n)}\right)^2 \qquad (15) \\
&\qquad \leqslant \frac{5(t-n+h(n)-1)^2}{h(n)}, \qquad (16)
\end{aligned}$$

where line (15) follows from bounding $t-l \leqslant t-n+h(n)-1$ and line (16) from observing that $\frac{t-n+h(n)-1}{h(n)} \geqslant 1$. Thus, we have:

$$\Pr\left(\left|\widehat{\mu}_i^{\text{R-ed},h(n)}(t) - \widetilde{\mu}_i^{\text{R-ed},h(n)}(t)\right| > \beta_i^{\text{R-ed},h(n)}(t,\delta_t)\right)$$

$$\leqslant \Pr\left(\left|\sum_{l=n-h(n)+1}^{n}\left(1+\frac{t-l}{h(n)}\right)X_l-\sum_{l=n-h(n)+1}^{n}\frac{t-l}{h(n)}\cdot X_{l-h(n)}\right|>h(n)\beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right)$$

$$2\exp\left(-\frac{\left(h(n)\beta_i^{\texttt{R-ed},h(n)}(t,\delta_t)\right)^2}{2\sigma^2\left(\frac{5(t-n+h(n)-1)^2}{h(n)}\right)}\right)=2\delta_t.$$

By replacing this result into Equation (14), and recalling the value of δ_t, we obtain:

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

□

Theorem 4.4. *Let $T\in\mathbb{N}$, then R-ed-UCB (Algorithm 1) with $B_i(t)\equiv\widehat{\mu}_i^{R\text{-}ed,h_{i,t}}(t)+\beta_i^{R\text{-}ed,h_{i,t}}(t)$, $h_{i,t}=\lfloor\epsilon N_{i,t-1}\rfloor$ 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:*

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

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:

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

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 $i^*\in\operatorname{argmax}_{i\in[K]}\left\{\sum_{l\in[T]}\mu_i(l)\right\}$. We decompose the above expression according to the good events $\mathcal{E}_t$:

$$R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB},T)=\sum_{t=1}^{T}\mathbb{E}[(\mu_{i*}(t)-\mu_{I_t}(N_{I_t,t}))\mathbb{1}\{\mathcal{E}_t\}]+\sum_{t=1}^{T}\mathbb{E}[(\mu_{i*}(t)-\mu_{I_t}(N_{I_t,t}))\mathbb{1}\{\neg\mathcal{E}_t\}]\tag{17}$$
$$\leqslant\sum_{t=1}^{T}\mathbb{E}[(\mu_{i*}(t)-\mu_{I_t}(N_{I_t,t}))\mathbb{1}\{\mathcal{E}_t\}]+\sum_{t=1}^{T}\mathbb{E}[\mathbb{1}\{\neg\mathcal{E}_t\}],\tag{18}$$

where we exploited $\mu_{i*}(t)-\mu_{I_t}(N_{I_t,t})\leqslant 1$ in line (18). Now, we bound the second summation, recalling that $\alpha>2$:

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

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

was bounded in Lemma A.4, we bound the summation with the integral as in Lemma C.4 to get:

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

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

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

Now, we work on the term inside the minimum:

$$\begin{aligned}B_{I_t}(t)-\mu_{I_t}(N_{I_t,t})&=\widehat{\mu}_{I_t}^{\text{R-ed},h_{I_t,t}}(t)+\beta_{I_t}^{\text{R-ed},h_{I_t,t}}(t,\delta_t)-\mu_{I_t}(N_{I_t,t}) &(19)\\&\leqslant\underbrace{\widetilde{\mu}_{I_t}^{\text{R-ed},h_{I_t,t}}(t)-\mu_{I_t}(N_{I_t,t})}_{\text{(a)}}+\underbrace{2\beta_{I_t}^{\text{R-ed},h_{I_t,t}}(t,\delta_t)}_{\text{(b)}}, &(20)\end{aligned}$$

where line (19) follows from the definition of $B_i(t)$, and line (20) 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. We start with (a):

$$\begin{aligned}\sum_{t=1}^{T}\min\left\{1,\widetilde{\mu}_{I_t}^{\text{R-ed},h_{I_t,t}}(t)-\mu_{I_t}(N_{I_t,t})\right\}&\leqslant 2K+\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\min\left\{1,\widetilde{\mu}_i^{\text{R-ed},h_{i,t_{i,j}}}(t_{i,j})-\mu_i(j)\right\}\\&\leqslant 2K+\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\min\left\{1,\frac{1}{2}(2t_{i,j}-2(j-1)+h_{i,t_{i,j}}-1)\gamma_i((j-1)-2h_{i,t_{i,j}}+1)\right\} &(21)\\&\leqslant 2K+\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\min\{1,T\gamma_i(j-2\lfloor\epsilon(j-1)\rfloor)\} &(22)\\&\leqslant 2K+\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\min\{1,T\gamma_i(\lfloor(1-2\epsilon)j\rfloor)\} &(23)\\&\leqslant 2K+T^q\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\gamma_i(\lfloor(1-2\epsilon)j\rfloor)^q &(24)\\&\leqslant 2K+T^q\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{i\in[K]}\sum_{j=\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor}\gamma_i(j) &(25)\\&\leqslant 2K+T^q\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{i\in[K]}\Upsilon_{\boldsymbol{\mu}}(\lfloor(1-2\epsilon)N_{i,T}\rfloor,q) &(26)\\&\leqslant 2K+KT^q\left\lceil\frac{1}{1-2\epsilon}\right\rceil\Upsilon_{\boldsymbol{\mu}}\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil,q\right), &(27)\end{aligned}$$

where line (21) follows from Lemma A.3, line (22) is obtained by bounding $2t_{i,j}-2(j-1)+h_{i,t_{i,j}}-1\leqslant 2T$ and exploiting the definition of $h_{i,t}=\lfloor\epsilon N_{i,t-1}\rfloor$, line (23) follows from the observation $j-2\lfloor\epsilon(j-1)\rfloor\geqslant j-2\epsilon(j-1)\geqslant\lfloor(1-2\epsilon)j\rfloor$, line (24) is obtained from the already exploited inequality $\min\{1,x\}\leqslant\min\{1,x\}^q\leqslant x^q$ for $q\in[0,1]$, line (25) is an application of Lemma C.1, line (26) applies the definition of $\Upsilon_{\boldsymbol{\mu}}(\cdot,q)$, and line (27) follows from Lemma C.2 recalling that $\sum_{i\in[K]}\lfloor(1-$

$2\epsilon)N_{i,T}\rfloor \leqslant (1-2\epsilon)T$.

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 $1+\lceil\frac{1}{\epsilon}\rceil$:

$$\sum_{t=1}^{T}\min\left\{1,2\beta_{I_t}^{\texttt{R-ed},h_{I_t,t}}(t,\delta_t)\right\} \leqslant K+K\left\lceil\frac{1}{\epsilon}\right\rceil+\sum_{i\in[K]}\sum_{j=\lceil\frac{1}{\epsilon}\rceil+1}^{N_{i,T}}\min\left\{1,2\sigma(t_{i,t}-(j-1)+h_{i,t_{i,t}}-1)\sqrt{\frac{10\log(t^\alpha)}{h_{i,t_{i,t}}^3}}\right\}.$$

$$=K+K\left\lceil\frac{1}{\epsilon}\right\rceil+\sum_{i\in[K]}\sum_{j=\lceil\frac{1}{\epsilon}\rceil+1}^{N_{i,T}}\min\left\{1,2\sigma T\sqrt{\frac{10\alpha\log(T)}{\lfloor\epsilon(j-1)\rfloor^3}}\right\}, \tag{28}$$

where line (28) follows from bounding $t^\alpha \leqslant T^\alpha$ 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:

$$2\sigma T\sqrt{\frac{10\alpha\log(T)}{\lfloor\epsilon(j-1)\rfloor^3}}\leqslant 1 \Longrightarrow \lfloor\epsilon(j-1)\rfloor \geqslant (2\sigma T)^{\frac{2}{3}}(10\alpha\log T)^{\frac{1}{3}}$$

$$\Longrightarrow j^*=\left\lceil\frac{1+\epsilon+(2\sigma T)^{\frac{2}{3}}(10\alpha\log T)^{\frac{1}{3}}}{\epsilon}\right\rceil.$$

Thus, we have:

$$K+K\left\lceil\frac{1}{\epsilon}\right\rceil+\sum_{i\in[K]}\sum_{j=\lceil\frac{1}{\epsilon}\rceil+1}^{N_{i,T}}\min\left\{1,2\sigma T\sqrt{\frac{10\alpha\log(T)}{\lfloor\epsilon(j-1)\rfloor^3}}\right\}\leqslant K+K\left\lceil\frac{1}{\epsilon}\right\rceil+\sum_{i\in[K]}\left(\sum_{j=\lceil\frac{1}{\epsilon}\rceil+1}^{j^*}1+\sum_{j=j^*+1}^{N_{i,T}}2\sigma T\sqrt{\frac{10\alpha\log(T)}{\lfloor\epsilon(j-1)\rfloor^3}}\right) \tag{29}$$

$$\leqslant K+K\left\lceil\frac{1}{\epsilon}\right\rceil+K\left(j^*-1-\left\lceil\frac{1}{\epsilon}\right\rceil+1\right)+2K\sigma T\sqrt{10\alpha\log(T)}\int_{x=j^*}^{+\infty}\frac{1}{(\epsilon(x-1)-1)^{\frac{3}{2}}}\mathrm{d}x \tag{30}$$

$$=K+Kj^*+\frac{4K\sigma T\sqrt{10\alpha\log(T)}}{\epsilon(\epsilon(j^*-1)-1)^{\frac{1}{2}}}$$

$$=K\left(3+\frac{1}{\epsilon}\right)+\frac{3K}{\epsilon}(2\sigma T)^{\frac{2}{3}}(10\alpha\log T)^{\frac{1}{3}}, \tag{31}$$

where line (29) is obtained by splitting the summation based on the value of j^*, line (30) comes from bounding the summation with the integral (Lemma C.4), and line (31) follows from substituting the value of j^* and simple algebraic manipulations. Putting all together, we obtain:

$$R_{\boldsymbol{\mu}}(\texttt{R-ed-UCB},T)\leqslant 1+\frac{2K}{\alpha-2}+5K+\frac{K}{\epsilon}+\frac{3K}{\epsilon}(2\sigma T)^{\frac{2}{3}}(10\alpha\log T)^{\frac{1}{3}}+KT^q\left\lceil\frac{1}{1-2\epsilon}\right\rceil\Upsilon_{\boldsymbol{\mu}}\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil,q\right)$$

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

□

A.2. Proofs of Section 5

Theorem 5.1 (Seznec et al., 2020)**.** *Let $\pi_{\boldsymbol{\mu}}^g$ be the* oracle greedy *policy:*

$$\pi_{\boldsymbol{\mu}}^g(t)\in\underset{i\in[K]}{\arg\max}\{\mu_i(t)\},\quad \forall t\in[T].$$

Then, $\pi^g_{\boldsymbol{\mu}}$ is optimal for the restless non-decreasing bandits (i.e., under Assumption 3.1).

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. □

Lemma A.5. *For every arm $i\in[K]$ and every round $t\in[T]$, let us define:*

$$\overline{\mu}_i^{R\text{-}less}(t) := \mu_i(t_{i,N_{i,t-1}}) + (t - t_{i,N_{i,t-1}}) \frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}},$$

if $N_{i,t-1}\geqslant 2$ else $\overline{\mu}_i^{R\text{-}less}(t) := +\infty$. Then, $\overline{\mu}_i^{R\text{-}less}(t)\geqslant\mu_i(t)$ and, if $N_{i,t-1}\geqslant 2$, it holds that:

$$\overline{\mu}_i^{R\text{-}less}(t) - \mu_i(t) \leqslant \big(t - t_{i,N_{i,t-1}}\big)\gamma_i(t_{i,N_{i,t-1}-1}).$$

Proof. Let us consider the following derivation:

$$\begin{aligned}
\mu_i(t) &= \mu_i(t_{i,N_{i,t-1}}) + \sum_{l=t_{i,N_{i,t-1}}}^{t-1} \gamma_i(l) \\
&\leqslant \mu_i(t_{i,N_{i,t-1}}) + (t - t_{i,N_{i,t-1}})\gamma_i(t_{i,N_{i,t-1}}) && (32)\\
&\leqslant \mu_i(t_{i,N_{i,t-1}}) + (t - t_{i,N_{i,t-1}}) \frac{\mu_i(t_{i,N_{i,t-1}}) - \mu_i(t_{i,N_{i,t-1}-1})}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}} =: \overline{\mu}_i^{\text{R-less}}(t), && (33)
\end{aligned}$$

where line (32) follows from Assumption 3.2 and line (33) from Lemma C.3. Moreover, if $N_{i,t-1}\geqslant 2$, we have:

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

where we employed Assumption 3.2 in the last line, noting $\frac{1}{t_{i,N_{i,t-1}} - t_{i,N_{i,t-1}-1}} \sum_{l=t_{i,N_{i,t-1}-1}}^{t_{i,N_{i,t-1}}-1} \gamma_i(l) \leqslant \gamma_i(t_{i,N_{i,t-1}-1})$. □

Theorem 5.2. *Let $T\in\mathbb{N}$, then `R-less-UCB` (Algorithm 1) with $B_i(t)\equiv\overline{\mu}_i^{R\text{-}less}(t)$ suffers an expected regret bounded, for every $q\in[0,1]$, as:*

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

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., $B_{i_t^*}(t) \leqslant B_{I_t}(t)$, we have:

$$\begin{aligned}\mu_{i*}(t) - \mu_{I_t}(t) + B_{I_t}(t) - B_{I_t}(t) &\leqslant \min\left\{1, \underbrace{\mu_{i_t^*}(t) - B_{i_t^*}(t)}_{\leqslant 0} + B_{I_t}(t) - \mu_{I_t}(t)\right\} \\ &\leqslant \min\{1, B_{I_t}(t) - \mu_{I_t}(t)\}. \end{aligned} \tag{34}$$

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

$$\begin{aligned} B_{I_t}(t) - \mu_{I_t}(t) &= \overline{\mu}_{I_t}^{\texttt{R-less}}(t) - \mu_{I_t}(t) && (35)\\ &\leqslant (t - t_{i,N_{i,t-1}})\gamma_i(t_{i,N_{i,t-1}-1}), && (36)\end{aligned}$$

where to get line (36) we applied Lemma A.5. Let us plug the expression derived in Equation (34) and decompose the summation of this term w.r.t. the K arms:

$$\begin{aligned} R_{\boldsymbol{\mu}}(\texttt{R-less-UCB}, T) &\leqslant \sum_{t=1}^{T} \min\left\{1, (t - t_{i,N_{i,t-1}})\gamma_i(t_{i,N_{i,t-1}-1}),\right\} \\ &= 2K + \sum_{i\in[K]} \sum_{j=3}^{N_{i,T}} \min\{1, (t_{i,j} - t_{i,j-1})\gamma_i(t_{i,j-2})\} \\ &\leqslant 2K + \sum_{i\in[K]} \sum_{j=3}^{N_{i,T}} (t_{i,j} - t_{i,j-1})^y \gamma_i(t_{i,j-2})^y && (37)\\ &\leqslant 2K + \sum_{i\in[K]} \left(\sum_{j=3}^{N_{i,T}} (t_{i,j} - t_{i,j-1})\right)^y \left(\sum_{j=3}^{N_{i,T}} \gamma_i(t_{i,j-2})^{\frac{y}{1-y}}\right)^{1-y} && (38)\\ &\leqslant 2K + T^y \sum_{i\in[K]} \left(\sum_{j=3}^{N_{i,T}} \gamma_i(j-2)^{\frac{y}{1-y}}\right)^{1-y} && (39)\\ &\leqslant 2K + T^y \sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(N_{i,T}, \frac{y}{1-y}\right)^{1-y} \\ &\leqslant 2K + T^y K^y \left(\sum_{i\in[K]} \Upsilon_{\boldsymbol{\mu}}\left(N_{i,T}, \frac{y}{1-y}\right)\right)^{1-y} && (40)\\ &\leqslant 2K + T^y K \Upsilon_{\boldsymbol{\mu}}\left(\left\lceil \frac{T}{K} \right\rceil, \frac{y}{1-y}\right)^{1-y}, && (41)\end{aligned}$$

line (37) follows from the inequality $\min\{1,x\} \leqslant \min\{1,x\}^y \leqslant x^y$ for $y \in \left[0, \frac{1}{2}\right]$, line (38) follows from Hölder's inequality with powers $\frac{1}{y} \geqslant 1$ and $\frac{1}{1-y} \geqslant 1$ (since $y \in \left[0, \frac{1}{2}\right]$), line (39) is obtained from observing that $\sum_{j=3}^{N_{i,T}} (t_{i,j} - t_{i,j-1}) \leqslant T$ and $\gamma_i(t_{i,j-2}) \leqslant \gamma_i(j-2)$ from Assumption 3.2, line (40) follows from Jensen's inequality as $y \in \left[0, \frac{1}{2}\right]$ and observing:

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

where line (41) is obtained from Lemma C.2. The final theorem statement is obtained by defining $q := \frac{y}{1-y} \in [0,1]$ and substituting it to the above equation.

□

Estimator Construction for the Stochastic Rising Restless Setting We provide the intuition behind the estimator construction and explain why it differs significantly from the one employed for the deterministic case. We start observing that for every $l \in \{2,\dots,N_{i,t-1}\}$, we have that:

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

where the inequality follows from Assumption 3.2. Since do not have access to $\gamma_i(t_{i,l-1})$ and we cannot directly estimate it, we need to perform a further bounding step. Specifically, based on Lemma C.3, we bound for every $l \in \{2,\dots,N_{i,t-1}\}$ and $h \in [l-1]$:

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

We report a first proposal of optimistic approximation of $\mu_i(t)$, i.e., $\widetilde{\widetilde{\mu}}_i^{\texttt{R-ed},h}(t)$, and the corresponding estimator, i.e., $\widehat{\widetilde{\mu}}_i^{\texttt{R-ed},h}(t)$, that are defined in terms of a window of size $1 \leqslant h \leqslant \lfloor N_{i,t-1}/2 \rfloor$:

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

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

Unfortunately, this estimator, 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:

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

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

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

The proof is composed of the following steps:

(i) Lemma A.6 shows that $\widetilde{\widetilde{\mu}}_i^{\texttt{R-less},h}(t)$ is an upper-bound for $\mu_i(t)$ and provides a bound to its bias for every value of h;

(ii) Lemma A.7 analyzes the concentration of $\widehat{\widetilde{\mu}}_i^{\texttt{R-less},h}(t)$ around $\widetilde{\widetilde{\mu}}_i^{\texttt{R-less},h}(t)$ for a specific choice of $\delta_t = t^{-\alpha}$ and when $h_{i,h} := h(N_{i,t-1})$ is a function of the number of pulls $N_{i,t-1}$ only;

(iii) Theorem 5.3 bounds the expected regret of `R-less-UCB` when $h_{i,h} = \lfloor \epsilon N_{i,t-1} \rfloor$ for $\epsilon \in (0,1/2)$.

Lemma A.6. *For every arm $i\in[K]$, every round $t\in[T]$, and window width $1\leqslant h\leqslant\lfloor N_{i,t-1}/2\rfloor$, 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, $\widetilde{\mu}_i^{\texttt{R-less},h}(t)\geqslant\mu_i(t_{i,N_{i,t-1}})$ and, if $N_{i,t-1}\geqslant 2$ it holds that:

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

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:

$$\begin{aligned}
\mu_i(t)&=\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l})+\sum_{j=t_{i,l}}^{t-1}\gamma_i(j)\right)\\
&\leqslant\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(\mu_i(t_{i,l})+(t-t_{i,l})\gamma_i(t_{i,l}-1)) &(42)\\
&\leqslant\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l})+\frac{t-t_{i,l}}{t_{i,l}-t_{i,l-h}}\sum_{j=t_{i,l-h}}^{t_{i,l}-1}\gamma_i(j)\right) &(43)\\
&=\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l})+(t-t_{i,l})\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{t_{i,l}-t_{i,l-h}}\right)\\
&\leqslant\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l})+(t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h}\right)=:\widetilde{\mu}_i^{\texttt{R-less},h}(t), &(44)
\end{aligned}$$

where lines (42) and (43) follow from Assumption 3.2, and line (44) is obtained from observing that $t_{i,l}\geqslant l$ and $t_{i,l}-t_{i,l-h}\geqslant h$. Concerning the bias, when $N_{i,t-1}\geqslant 2$, we have:

$$\begin{aligned}
\widetilde{\mu}_i^{\texttt{R-less},h}(t)-\mu_i(t)&=\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}\left(\mu_i(t_{i,l})+(t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h}\right)-\mu_i(t)\\
&\leqslant\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{h} &(45)\\
&=\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{t_{i,l}-t_{i,l-h}}\cdot\frac{t_{i,l}-t_{i,l-h}}{h}\\
&\leqslant\frac{1}{h}\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l)\gamma_i(t_{i,l-h})\cdot\frac{t_{i,l}-t_{i,l-h}}{h} &(46)\\
&\leqslant\frac{t_{i,N_{i,t-1}}-t_{i,N_{i,t-1}-2h+1}}{h^2}\gamma_i(t_{i,N_{i,t-1}-2h+1})\sum_{l=N_{i,t-1}-h+1}^{N_{i,t-1}}(t-l) &(47)
\end{aligned}$$

$$= \frac{(2t-2N_{i,t-1}+h-1)(t_{i,N_{i,t-1}}-t_{i,N_{i,t-1}-2h+1})}{2h}\gamma_i(t_{i,N_{i,t-1}-2h+1}), \tag{48}$$

where line (45) follows from observing that $\mu_i(t_{i,l}) \leqslant \mu_i(t)$, line (46) derives from Assumption 3.2 and bounding $\frac{\mu_i(t_{i,l})-\mu_i(t_{i,l-h})}{t_{i,l}-t_{i,l-h}} \leqslant \gamma_i(t_{i,l-h})$, line (47) is obtained by bounding $t_{i,l}-t_{i,l-h} \leqslant t_{i,N_{i,t-1}}-t_{i,N_{i,t-1}-2h+1}$ and $\gamma_i(t_{i,l-h}) \leqslant \gamma_i(t_{i,N_{i,t-1}-2h+1})$, and line (48) follows from computing the summation. □

Lemma A.7. *For every arm $i \in [K]$, every round $t \in [T]$, and window width $1 \leqslant h \leqslant \lfloor N_{i,t-1}/2 \rfloor$, let us define:*

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

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

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

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

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 $\{h_{i,t} \geqslant 1\}$. Hence:

$$\begin{aligned}
&\Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h_{i,t}}(t) - \widetilde{\mu}_i^{\texttt{R-less},h_{i,t}}(t) \right| > \beta_i^{\texttt{R-less},h_{i,t}}(t,\delta_t) \right) && (49)\\
&\quad = \Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h(N_{i,t-1})}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(N_{i,t-1})}(t) \right| > \beta_i^{\texttt{R-less},h(N_{i,t-1})}(t,\delta_t) \right) && (50)\\
&\quad \leqslant \Pr\left(\exists n \in \{0,\dots,t-1\} \text{ s.t. } h(n) \geqslant 1 : \left| \widehat{\mu}_i^{\texttt{R-less},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(n)}(t) \right| > \beta_i^{\texttt{R-less},h(n)}(t,\delta_t) \right)\\
&\quad \leqslant \sum_{n \in \{0,\dots,t-1\} : h(n) \geqslant 1} \Pr\left(\left| \widehat{\mu}_i^{\texttt{R-less},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(n)}(t) \right| > \beta_i^{\texttt{R-less},h(n)}(t,\delta_t) \right), && (51)
\end{aligned}$$

where line (50) follows from the definition of $h_{i,t} = h(N_{i,t-1})$, and line (51) 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 the case. 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:

$$h(n)\left(\widehat{\mu}_i^{\texttt{R-less},h(n)}(t) - \widetilde{\mu}_i^{\texttt{R-less},h(n)}(t) \right) = \sum_{l=n-h(n)+1}^{n} \left(X_{t_{i,l}} + (t-l)\frac{X_{t_{i,l}} - X_{t_{i,l-h(n)}}}{h(n)} \right) = \sum_{s=1}^{t-1} \epsilon_s Y_s X_s, \tag{52}$$

where:

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

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:

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

and we introduce the random variables $\widetilde{X}_j:=X_{t_{i,j}}$ and $\widetilde{Y}_j:=Y_{t_{i,j}}$. 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 (Doob, 1953; Bubeck et al., 2008). We introduce the filtration $\mathcal{F}_{\tau-1}=\sigma(I_1,R_1,\dots,I_{\tau-1},R_{\tau-1},I_\tau)$ 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:

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

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

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

From which, by substituting into Equation (52) 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 A.4:

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

Thus, we can now apply Azuma-Höeffding's inequality (Lemma C.5):

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

By replacing into Equation (51) and summing over n, we obtain:

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

□

Theorem 5.3. *Let $T\in\mathbb{N}$, then* `R-less-UCB` *(Algorithm 1) 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:*

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

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:

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

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 $\mathcal{E}_t$:

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

where we exploited $\mu_{i_t^*}(t) - \mu_{I_t}(t) \leqslant 1$ in the inequality. Now, we bound the second summation, as done in Theorem 4.4:

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

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., $B_{i_t^*}(t) \leqslant B_{I_t}(t)$:

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

Now, we work on the term inside the minimum:

$$\begin{aligned} B_{I_t}(t) - \mu_{I_t}(t) &= \widehat{\mu}_{I_t}^{\texttt{R-less}, h_{I_t,t}}(t) + \beta_{I_t}^{\texttt{R-less}, h_{I_t,t}}(t) - \mu_{I_t}(t) && (53) \\ &\leqslant \underbrace{\widetilde{\mu}_{I_t}^{\texttt{R-less}, h_{I_t,t}}(t) - \mu_{I_t}(t)}_{\text{(a)}} + \underbrace{2\beta_{I_t}^{\texttt{R-less}, h_{I_t,t}}(t)}_{\text{(b)}}, && (54) \end{aligned}$$

where line (53) follows from the definition of $B_i(t)$ and line (54) from the good event $\mathcal{E}_t$. We proceed decomposing over the

arms, starting with (a):

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

$$\leqslant 2K+\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\min\left\{1,\frac{(2t_{i,j}-2(j-1)+h_{i,t_{i,j}}-1)(t_{i,j-1}-t_{i,j-2h_{i,t}+1})}{2h_{i,t}}\gamma_i(t_{i,(j-1)-2h_{i,t_{i,j}}+1})\right\} \tag{55}$$

$$\leqslant 2K+\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\min\left\{1,\frac{T^2}{\lfloor\epsilon(j-1)\rfloor}\gamma_i(t_{i,j-2\lfloor\epsilon(j-1)\rfloor})\right\} \tag{56}$$

$$\leqslant 2K+\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\min\left\{1,\frac{T^2}{\lfloor\epsilon(j-1)\rfloor}\gamma_i(\lfloor(1-2\epsilon)j\rfloor)\right\} \tag{57}$$

$$\leqslant 2K+T^{2z}\sum_{i\in[K]}\sum_{j=3}^{N_{i,T}}\left(\frac{\gamma_i(\lfloor(1-2\epsilon)j\rfloor)}{\lfloor\epsilon(j-1)\rfloor}\right)^{z} \tag{58}$$

$$\leqslant 2K+T^{2z}\sum_{i\in[K]}\left(\sum_{j=3}^{N_{i,T}}\frac{1}{\lfloor\epsilon(j-1)\rfloor}\right)^{z}\left(\sum_{j=3}^{N_{i,T}}\gamma_i(\lfloor(1-2\epsilon)j\rfloor)^{\frac{z}{1-z}}\right)^{1-z} \tag{59}$$

$$\leqslant 2K+T^{2z}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{i\in[K]}\left(\sum_{j=\lfloor 2\epsilon\rfloor}^{\lfloor\epsilon(N_{i,T}-1)\rfloor}\frac{1}{j}\right)^{z}\left(\sum_{j=\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor}\gamma_i(j)^{\frac{z}{1-z}}\right)^{1-z} \tag{60}$$

$$\leqslant 2K+T^{2z}(1+\log(\epsilon T))^{z}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil\sum_{i\in[K]}\left(\sum_{j=\lfloor 3(1-2\epsilon)\rfloor}^{\lfloor(1-2\epsilon)N_{i,T}\rfloor}\gamma_i(j)^{\frac{z}{1-z}}\right)^{1-z} \tag{61}$$

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

$$\leqslant 2K+T^{2z}(1+\log(\epsilon T))^{z}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil K^{z}\left(\sum_{i\in[K]}\Upsilon_{\boldsymbol{\mu}}\left(\lfloor(1-2\epsilon)N_{i,T}\rfloor,\frac{z}{1-z}\right)\right)^{1-z} \tag{62}$$

$$\leqslant 2K+T^{2z}(1+\log(\epsilon T))^{z}\left\lceil\frac{1}{\epsilon}\right\rceil\left\lceil\frac{1}{1-2\epsilon}\right\rceil K\Upsilon_{\boldsymbol{\mu}}\left(\left\lceil(1-2\epsilon)\frac{T}{K}\right\rceil,\frac{z}{1-z}\right)^{1-z}. \tag{63}$$

where line (55) follows from the bias bound of Lemma A.6, line (56) is obtained from bounding $(2t_{i,j}-2(j-1)+h_{i,t_{i,j}}-1)(t_{i,j-1}-t_{i,j-2h_{i,t}+1})\leqslant 2T^2$ and using the definition of $h_{i,t}$, line (57) derives from observing that $\gamma_i(t_{i,j})\leqslant\gamma_i(j)$ for Assumption 3.2 and having bounded the floor analogously as done in Theorem 4.4, line (58) from the inequality $\min\{1,x\}\leqslant\min\{1,x\}^z\leqslant x^z$ for $z\in[0,1/2]$, line (59) is obtained from Hölder's inequality with exponents $\frac{1}{z}\geqslant 1$ and $\frac{1}{1-z}\geqslant 1$ respectively, line (60) is an application of Lemma C.1 to independently to both inner summations, line (61) derives from bounding the harmonic sum, i.e., $\sum_{\lfloor 2\epsilon\rfloor}^{\lfloor\epsilon(N_{i,T}-1)\rfloor}\frac{1}{j}\leqslant 1+\log(\epsilon(N_{i,T}-1))\leqslant 1+\log(\epsilon T)$, line (62) follows from Jensen's inequality, line (63) is obtained from Lemma C.2. 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:

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

Putting all together, we obtain:

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

□

B. Bounding the Cumulative Increment

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

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

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. Rested Setting

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

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

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

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

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

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

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

B.2. Restless Setting

Let us now move to the restless setting. From Theorem 5.2, we have:

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

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

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

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

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:

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

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

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

C. Technical Lemmas

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

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

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 C.2. *Under Assumption 3.2, it holds that:*

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

Proof. We first claim that there exists an optimal assignment of $N^*_{i,T}$ are such that $|N^*_{i,T} - N^*_{i',T}| \leqslant 1$ 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:

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

where the inequality follows from Assumption 3.2. 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 $|\widetilde{N}^*_{i_1,T} - \widetilde{N}^*_{i_2,T}| \leqslant 1$. Thus, we have found a better solution to the optimization problem, contradicting the hypothesis. Since the optimal assignment fulfills $|N^*_{i,T} - N^*_{i',T}| \leqslant 1$, it must be that $N^*_{i,T} \leqslant \left\lceil \frac{T}{K} \right\rceil$ for all $i \in [K]$. □

Lemma C.3. *Under Assumptions 3.1 and 3.2, for every $i\in[K]$, $k,k'\in\mathbb{N}$ with $k'<k$, for both rested and restless bandits, it holds that:*

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

Proof. Using Assumption 3.2, we have:

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

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

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

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

If f is monotonically non-increasing, then:

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

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 $f(x) \geqslant f(x_{i-1})$ and consequently $\int_{I_i} f(x)\mathrm{d}x \geqslant f(x_{i-1})\mathrm{vol}(I_i) = f(x_{i-1})$. Thus:

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

Recalling that $f(b) \leqslant \int_{x=b}^{b+1} f(x)\mathrm{d}x$, 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 $f(x) \geqslant f(x_i)$ and consequently $\int_{I_i} f(x)\mathrm{d}x \geqslant f(x_i)$. Thus:

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

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

Theorem C.5 (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 $\mathbb{E}[\exp(\lambda X_t)|\mathcal{F}_{t-1}] \leqslant \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ for any $\lambda > 0$ (i.e., σ^2-subgaussian). Let $\alpha_1, \ldots, \alpha_n$ be non-negative real numbers. Then, for every $\kappa \geqslant 0$ it holds that:*

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

Proof. It is a straightforward extension of Azuma-Höeffding inequality for subgaussian random variables. We apply the Chernoff's method for some $s > 0$:

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

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:

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

An inductive argument, leads to:

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

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

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

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 C.6. *Let $\Upsilon_{\boldsymbol{\mu}}(T,q)$ be as defined in Equation (2) for some $q\in[0,1]$. Then, for all $i\in[K]$ and $l\in\mathbb{N}$ the following statements hold:*

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

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 C.4. Let us start with $\gamma_i(l)\leqslant be^{-cl}$:

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

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

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

For $cq=1$, we obtain:

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

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

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

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

D. Efficient Update

Under the assumption that the window size depends on the number of pulls only and that $0\leqslant h(n+1)-h(n)\leqslant 1$, 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 $r_i(n) := R_{t_{i,n}}$.

E. Experimental Setting and Additional Results

E.1. Parameter Setting

The choices of the parameters of the 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. (2014);
- `KL-UCB`: $c = 3$ as required by the theoretical results on the regret provided by Garivier & Cappé (2011);
- `Ser4`: according to what suggested by Allesiardo et al. (2017) we selected $\delta = 1/T$, $\epsilon = \frac{1}{KT}$, and $\phi = \sqrt{\frac{N}{TK \log(KT)}}$;
- `SW-UCB`: as suggested by Garivier & Moulines (2011) we selected the sliding-window $\tau = 4\sqrt{T \log T}$ and the constant $\xi = 0.6$;
- `SW-KL-UCB` as suggested by Garivier & Moulines (2011) we selected the sliding-window $\tau = \sigma^{-4/5}$;
- `SW-TS`: as suggested by Trovò et al. (2020) for the smoothly changing environment we set $\beta = 1/2$ and sliding-window $\tau = T^{1-\beta} = \sqrt{T}$.

E.2. IMDB Experiment

We created a bandit environment in which each of the classification algorithms is an arm of the bandit. The interaction for each round $t \in T$ of the real-world experiment is composed by the following:

- the agent decides to pull arm I_t;
- a random point x_t of the IMDB dataset is selected and supplied to the classification algorithm associated to arm I_t;
- the "base" algorithm classifies the sample, i.e., it 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 $R_t = 1 - |y_t - \hat{y}_t|$;
- the base algorithm is updated using (x_t, y_t);

Since the base algorithms are trained only if their arm is selected, this is a problem which belongs to the rested scenario.

For the classification task we decided to employ:

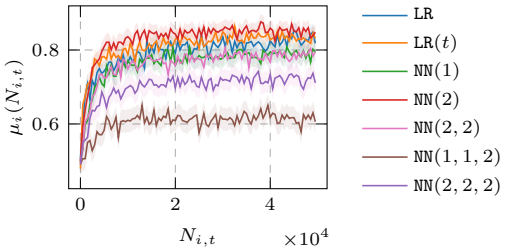


Figure 6. Empirical learning curves of the classification algorithms (arms) of the IMDB experiment

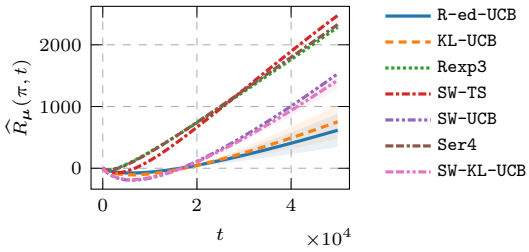


Figure 7. Cumulative regret in a 2-arms online model selection on IMDB dataset (30 runs, 95% c.i.).

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

Specifically, we adopt a decreasing scheme for the learning rate of $\lambda_t = \frac{\beta}{t}$ (denoted with LR(t) from now on) and a constant learning rate $\lambda_t = \beta$ (denoted as LR from now on). Moreover, the NNs use as activation functions the rectified linear unit, i.e., $relu(x) = \max(0,x)$, a constant learning rate $\alpha = 0.001$ and the "adam" stochastic gradient optimizer for fitting. 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 analyzed their global performance on the IMDB dataset by averaging $1{,}000$ independent runs in which each strategy is sequentially fed with all the available $50{,}000$ samples. The goal was to determine, at each step, the value of the payoff $\mu_i(n)$. Figure 6a provides the average learning curves of the selected algorithms. As we expected, from a qualitative perspective, the average learning curves are increasing and concave, however, due to the limited number of simulations, Assumptions 3.1 and 3.2 are not globally satisfied.

We also perform an experiment using only LR(t) and LR as arms. Figure 7a reports the result of a run of the MAB algorithms over the IMDB scenario. The analogy between this result and the one of the 2-arms synthetic rested bandit (Figure 4b) is clear, indeed R-ed-UCB outperforms the other baselines when the learning curves of the base algorithms at some point intersects one another.

E.3. Pulls of each arm

Figure 8 presents the average number of pulls for each arm for each one of the algorithm analysed in the synthetic experiments of Section 6. Figure 8a shows how R-less-UCB is able to identify and discard the majority of the suboptimal arms using a few pulls, and it is second only to SW-TS, which seems to commit to a single arm which turns out to be the optimal one (arm 13). Figure 8b shows that R-ed-UCB explored arms 13 and 1 more than the others, which are respectively the best and the second-best, and most likely needs a longer time horizon to select which one is the best among the twos. Figure 8c highlights the fact that R-ed-UCB undoubtedly identified which arm is the best (arm 2), while KL-UCB, SW-UCB, SW-KL-UCB do not identify the best arm. Ser4, Rexp3 and SW-TS pulled the best arm slightly more than 50% of the times, paying the already discussed initial learning phase.

E.4. Additional Experimental Results

We evaluated the performance of the algorithms over 50 different bandits with $K \in \{2,\dots,15\}$ randomly generated arms over a time horizon of $T = 200{,}000$ rounds. We averaged the run of each algorithm on a single scenario over 10 independent experiments and compared the expected value of the ranking of the considered algorithms in order to draw up a leaderboard: in every scenario we ranked the algorithms based on their empirical regret, giving the first placement to the one with the lowest value. We report the summarized results of the rank of the algorithms averaged over the 50 experiments in Table 2. In the rested case R-ed-UCB is among the worse ones (4.98 on average), however this is again due to the fact that on average the algorithm is not superior to the baselines, which conversely do not provide any theoretical guarantees in some

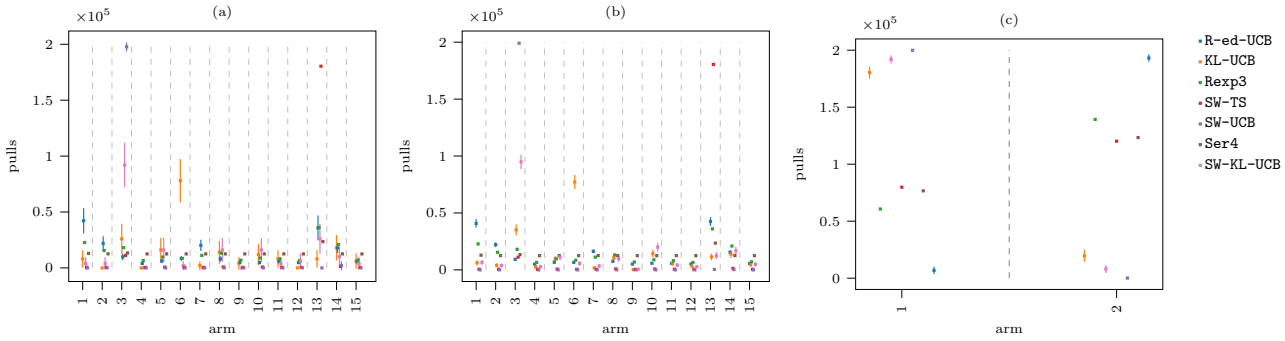


Figure 8. Average number of pulls: (a) 15 arms `R-less`, (b) 15 arms `R-ed`, (c) 2 arms `R-ed`.

Table 2. Ranking of the algorithms (50 bandits, 10 runs, 95% c.i. in brackets).

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

specific settings (see the 2 arm experiment). In the restless case `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 the `R-less-UCB` upper bound as follows:

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

which we call `R-less-UCB-H` to denote it is an heuristic method, i.e., not having theoretical results on the regret. 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.