A single algorithm for both restless and rested rotting bandits

Julien Seznec
Lelivrescolaire.fr
SCOOL, Inria Lille

Pierre Menard
SCOOL, Inria Lille

Alessandro Lazaric
FAIR Paris

Michal Valko
DeepMind Paris

Abstract

In many application domains (e.g., recommender systems, intelligent tutoring systems), the rewards associated to the actions tend to decrease over time. This decay is either caused by the actions executed in the past (e.g., a user may get bored when songs of the same genre are recommended over and over) or by an external factor (e.g., content becomes outdated). These two situations can be modeled as specific instances of the rested and restless bandit settings, where arms are *rotting* (i.e., their value decrease over time). These problems were thought to be significantly different, since Levine et al. (2017) showed that state-of-the-art algorithms for restless bandit perform poorly in the rested rotting setting. In this paper, we introduce a novel algorithm, Rotting Adaptive Window UCB (`RAW-UCB`), that achieves near-optimal regret in both rotting rested and restless bandit, without any prior knowledge of the setting (rested or restless) and the type of non-stationarity (e.g., piece-wise constant, bounded variation). This is in striking contrast with previous negative results showing that no algorithm can achieve similar results as soon as rewards are allowed to increase. We confirm our theoretical findings on a number of synthetic and dataset-based experiments.

1 Introduction

When we design sequential learner, we would like them to be as adaptive to environment as possible. This becomes a challenge when the environment only provides limited feedback, as in the *bandit* setting (Lai and Robbins, 1985; Lattimore and Szepesvári, 2020), where the learner receives only the feedback associated to the action it executed. Since the early stages of the research in bandits (Thompson, 1933; Whittle, 1980), one of the most desirable properties for a learners would be to adapt to actions whose *value changes over time* (Whittle, 1988), as it happens in non-stationary environments. In fact, from applications in medical trials (where the patient can become more resistant to antibiotics) to a modern applications in recommender systems (Chapelle and Li, 2011; Tracà and Rudin, 2015), assuming that the environment is *stationary is very limiting.*

However, modeling and managing non-stationary environments is obviously way more difficult (Lattimore and Szepesvári, 2020). That is why Auer et al. (2003) went as far as to consider the worst-case scenario, referred to as the *adversarial bandit* setting, where the learner should try to shield from the worst possible variation in rewards. Nonetheless, real-world environments are rarely adversarial and algorithms for adversarial bandits turn out to be too conservative for practical use. On the one hand, in order to manage such general family of environments, the performance of a learner is compared to the best *fixed* action in *hindsight*. This is arguably a weaker objective w.r.t. competing against the optimal strategy, as it is the case in stationary bandits. On the other hand, state-of-the-art adversarial algorithms (Audibert and Bubeck, 2009), which are proved to recover near-optimal regret rates on stationary problems, still under-perform in practice against optimal stationary algorithm (Zimmert and Seldin, 2019). In order to address these issues, prior work identified specific types of non-stationary environments, for which specifically designed algorithms can be used.

There are two main classes of non-stationary environments, depending on whether the change of rewards is triggered by the actions of the learner, the *rested bandits*, or it happens over time independently from the learner, the *restless bandits.* In this paper, we consider the specific case where the changes in the rewards are

Proceedings of the 23^{rd}International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

arbitrary *non-increasing* functions of time and/or number of pulls (in contrast with typical restless bandit models, where the evolution of rewards was regulated by Markov chain processes). For instance, Warlop et al. (2018) model boredom effects in recommender systems as a rested bandit problem, but need to resort to a more general reinforcement learning framework to address the fact that rewards are decreasing while an action is repeatedly selected but may increase back if *enough time* has passed since the last time is chosen. Immorlica and Kleinberg (2018) and Pike-Burke and Grunewalder (2019) have recently modeled these recharging effects as a bandits problem. In the restless setting, Louëdec et al. (2016) models obsolescence of appearing arms (e.g. piece of news) with a known exponential rate. Komiyama and Qin (2014) study a parametric decay in restless bandits where rewards are linear combination of known decaying function. In the following, we briefly review the most relevant results available for restless bandit (where no rotting assumption has been studied before) and the rested rotting bandit settings.

Restless stochastic bandits Garivier and Moulines (2011) study the restless bandits case, where rewards are piece-wise stationary. If the number of stationary pieces Υ_T at the horizon T is known, the optimal strategy is included in a set of T^{Υ_T} switching experts. Hence one can use Exp3.S, an adversarial algorithm designed for this specific set of experts (Auer et al., 2003). Moreover, Garivier and Moulines (2011) show that two upper-confidence bound index algorithms with passive forgetting parameters, SW-UCB and D-UCB, are also able to reach nearly-minimax performance when they know in advance Υ_T and T. Recent research (Cao et al., 2019; Liu et al., 2018; Besson and Kaufmann, 2019) has focused on integrating change-detection algorithms with standard bandit learners (*e.g.* UCB) to actively forget past rewards whenever a significant variation in the reward distribution is detected. Among them, we mention GLR-klUCB (Besson and Kaufmann, 2019) which uses a parameter-free change-point detector. These algorithms actively explore sub-optimal actions to track potential increase in their value. Yet, their analysis assume that change-points are always big enough to be detectable with high-probability. Auer et al. (2019) introduce AdSwitch, a filtering algorithm with a planned active exploration scheme for sub-optimal actions. AdSwitch achieves the minimax rate while being agnostic to Υ_T without any extra assumption.

Besbes et al. (2014) introduced a restless bandits framework where the environment has a variation budget of V_T to change the rewards' values. In this setup, the best arm can change at each round and thus the optimal strategy is not necessary included in a "small" set of switching experts. Yet, they show that the best strategy with $\mathcal{O}\left(T^{1/3}\right)$ switches suffers low regret compared to the optimal strategy. Hence, Exp3.S matches the minimax rate $\mathcal{O}\left(T^{2/3}\right)$ with the knowledge of V_T. Cheung et al. (2019) and Russac et al. (2019) extended SW-UCB and D-UCB to show that they also match the minimax rate of the variation budget setting even in the more general linear bandits framework. Chen et al. (2019) show that AdSwitch also matches the minimax rate without the knowledge of V_T. They also analyse ADA-ILTCB+, an algorithm which achieves similar guarantee in the more general linear setting. Wei et al. (2016) extended these results to a non-stationary environment where both the means and the variances of the rewards may change.

Rested rotting bandits Finally, Heidari et al. (2016); Levine et al. (2017) and Seznec et al. (2019) studied *rested rotting bandits*, when the reward of an action decreases every time it is pulled. Seznec et al. (2019) recently proposed a nearly-optimal algorithm for this setting. Interestingly, the algorithm does not execute an *index policy* (defined later) which is a prevalent choice in bandit. Actually, a previous attempt of using an index policy by Levine et al. (2017) resulted in a sub-optimal performance.

Our contribution is threefold:

- We show that no learning strategy can achieve $o(T)$ worst case rate when we allow for both rested and restless decay (Section 2).
- We introduce a novel index policy RAW-UCB (Section 3) and prove that it achieves minimax rate regret for either restless (Section 4) or rested (Section 5) settings without any prior knowledge of the type of decay, the amount of change, or the horizon.
- RAW-UCB also recovers problem-dependent $\mathcal{O}(\log T)$ bounds in both setups. In the restless case[1], such bounds cannot be achieved when the reward can increase. Hence, it shows that the decreasing assumption do help the learner compared to the well-studied general case.

Also, we provide a rested simulated (Appendix G.1) and restless real-world (Section 6) benchmarks on which RAW-UCB gives the most consistent results in both setups.

[1] In the rested case, Heidari et al. (2016) shows that increasing reward is a much harder problem, even in the absence of noise.

2 Decreasing multi-armed bandits

At each round t, an agent chooses an arm $i_t \in \mathcal{K} \triangleq \{1, ..., K\}$ and receives a noisy reward o_t. The sample associated to each arm i is a σ^2-sub-Gaussian r.v. with expected value of $\mu_i(t, n)$ which depends on the number of times n it was pulled before and on the time t.

Let $\mathbf{H}_t \triangleq \{\{i(s), o_s\}, \forall s \leq t\}$ be the sequence of arms pulled and rewards observed until round t, then

$$o_t \triangleq \mu_{i_t}(t, N_{i_t, t-1}) + \varepsilon_t \,,$$

with $\mathbb{E}\left[\varepsilon_t | \mathbf{H}_{t-1}\right] = 0$ and $\forall \lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda \varepsilon_t}\right] \leq e^{\frac{\sigma \lambda^2}{2}}$, where $N_{i,t} \triangleq \sum_{s=1}^{t} \mathbb{1}\left(i_s = i\right)$ is the number of pulls of arm i at time t. We call $\mu \triangleq \{\mu_i\}_{i \in \mathcal{K}}$ the set of reward functions.

Decreasing rewards Throughout all the paper, we consider the following assumption.

Assumption 1. *For each arm i, any number of pulls n, and time t, the functions $\mu_i(t, \cdot)$ and $\mu_i(\cdot, n)$ are non-increasing.*

We will use interchangeably the terms *decreasing*, *decaying* and *rotting* to refer to this Assumption. If $\mu_i(t, N_{i,t}) = \mu_i(N_{i,t})$, then i is called a rested arm. If $\mu_i(t, N_{i,t}) = \mu_i(t)$, then i is called a restless arm.

Learning problem A (deterministic) learning policy π is a function that maps history of observations to arms, i.e., $\pi(\mathcal{H}_t) \in \mathcal{K}$. In the following, we often use $\pi(t) \triangleq \pi(\mathcal{H}_{t-1})$ to denote the arm pulled at time t. The performance of a policy π is measured by the (expected) rewards accumulated over time,

$$J_T(\pi, \mu) \triangleq \sum_{t=1}^{T} \mu_{\pi(t)}\left(t, N_{\pi(t), t-1}\right) \,.$$

A (deterministic) oracle policy is a function which maps the set of reward functions and a round to an arm, i.e., $\pi(t, \mu) \in \mathcal{K}$. Thus, these oracles have access to the true (without noise) value of the rewards, including future value. Notice that at the horizon T, there are K^T distinct deterministic policies. Therefore, we call an optimal (oracle) policy, one which, at a given horizon T, maximizes the reward

$$\pi_T^*(t, \mu) \in \arg\max_{\pi \in \mathcal{K}^T} J_T(\pi, \mu) \,.$$

We define the regret as

$$R_T(\pi, \mu) \triangleq J_T(\pi_T^\star, \mu) - J_T(\pi, \mu).$$

Notice that this definition is more challenging than the regret w.r.t. the best fixed-arm policy commonly used as comparator in adversarial bandits. In the following, we often use shorter notation $\pi_T^*(t)$, $J_T(\pi)$, $R_T(\pi)$ where the considered problem μ is implicit.

Greedy oracle policy It is still unclear if 1) we can compute $\pi_T^\star$ in a tractable way; 2) if a learning policy can suffer low regret compared to this policy. We call π_O the oracle policy which selects greedily at each round t the largest available reward $i_t \in \arg\max_{i \in \mathcal{K}} \mu_i(t, N_{i,t-1})$.[2] We notice that this policy is optimal at any time in any restless non-stationary bandit problem $\mu(t)$. Heidari et al. (2016) show that it is also optimal in the rested rotting bandits problem. Thus, π_O answers positively to the first question for either rested or restless decay. In the next proposition, we show that the greedy oracle suffers linear worst-case regret when we allow for both restless and rested decay at the same time. Worse, we show that no learning policy can approach the performance of the optimal oracle at a $o(T)$ rate

Proposition 1. *In the no noise setting ($\sigma = 0$), there exists a rotting 2-arms bandits problem (satisfying Assumption 1) with reward value in $[0, 1]$, with one rested arm and one restless arm, and with at most one change-point before T each, such that the greedy oracle strategy π_O suffers a regret*

$$R_T\left(\pi_O\right) \geq \left\lfloor \frac{T}{4} \right\rfloor .$$

Moreover, for any learning strategy π_S, there exists a rotting 2-arms bandits problem (satisfying Assumption 1) with reward value in $[0, 1]$, with one rested arm and one restless arm, and with at most one change-point before T each, such that

$$R_T\left(\pi_S\right) \geq \left\lfloor \frac{T}{8} \right\rfloor .$$

Notice that the two reward functions of the constructed difficult problems are simple: either rested or restless, bounded and with at most one break-point. If we consider a 2-arm setup with one rested arm and one restless arm, a good strategy may be to select the restless arm even when its current value is the worst. Indeed, this value is only available now, while the good value of the rested arm will still be available in the future. Whether the restless rewards are interesting to the learner depends on the future behavior of the (currently best) rested arm. On the first hand, if it decays below the current value of the restless arm before the horizon T, then the learner should profit from the restless reward available right now. On the other hand, if the rested arm stays optimal until the end of the game then the learner should ignore the restless arm and follows the greedy oracle strategy. However, the learner does not know in advance if (and how much) an arm will decay and any anticipation she makes will

[2] We break the ties arbitrarily, for instance by selecting the smallest index in $\arg\max_{i \in \mathcal{K}} \mu_i\left(t, \mathcal{H}_t\right)$

turn to be bad in the worst case. We formalize these ideas in the proof in Appendix B and show that any strategy suffers linear regret in the worst case.

While learning with rested and restless rotting reward is impossible, we show in the next sections that a single policy reaches near-optimal guarantee in both separated setups.

3 The RAW-UCB algorithm

Notation For policy π, we define the average of the last h observations of arm i at time t as

$$\widehat{\mu}_i^h(t,\pi) \triangleq \frac{1}{h}\sum_{s=1}^{t-1} \mathbb{1}(\pi(s) = i \wedge N_{i,s} > N_{i,t-1} - h)\, o_s \quad (1)$$

and the average of the associated means as

$$\overline{\mu}_i^h(t,\pi) \triangleq \frac{1}{h}\sum_{s=1}^{t-1} \mathbb{1}(\pi(s) = i \wedge N_{i,s} > N_{i,t-1} - h)\, \mu_i(s, N_{i,s-1})\,.$$

A favorable event We use a similar high probability analysis than UCB1. We design a favorable event and we show in Prop. 2 that it holds with high probability.

Proposition 2. *For any round t and confidence $\delta_t \triangleq 2t^{-\alpha}$, let*

$$\xi_t^\alpha \triangleq \Big\{\forall i \in \mathcal{K},\ \forall n \le t-1,\ \forall h \le n, \\ |\widehat{\mu}_i^h(t,\pi) - \overline{\mu}_i^h(t,\pi)| \le c(h,\delta_t)\Big\} \quad (2)$$

be the event under which the estimates at round t are all accurate up to $c(h,\delta_t) \triangleq \sqrt{2\sigma^2\log(2/\delta_t)/h}$. Then, for a policy π which pulls each arms once at the beginning, and for all $t > K$,

$$\mathbb{P}\Big[\overline{\xi_t^\alpha}\Big] \le \frac{Kt^2\delta_t}{2} = Kt^{2-\alpha}\,.$$

Rotting Adaptive Window Upper Confidence Bound (RAW-UCB or π_{R}). At each round, RAW-UCB selects the arm with the largest following index,

$$\operatorname{ind}(i,t,\delta_t) \triangleq \min_{h \le N_{i,t-1}} \widehat{\mu}_i^h(t,\pi_{\mathrm{R}}) + c(h,\delta_t), \quad (3)$$

with $\delta_t \triangleq \frac{2}{t^\alpha}$. There is a bias-variance trade-off for the window choice: more variance for smaller size of the window h and more bias for larger h. The goal of RAW-UCB is to adaptively select the right window to compute the tightest UCB. RAW-UCB uses the indexes of UCB1 computed on all the slices of each arm's history which include the last pull. When the rewards are rotting, all these indexes are upper confidence bounds on the *next value*. Thus, RAW-UCB simply selects the tightest (minimum) one as index of the arm: it is a pure UCB-index algorithm. By contrast, when reward can increase, the learner can only derive upper-confidence bound on past values which are loosely related to the next value. Hence, all the UCB-index algorithms in the restless non-stationary literature need to add change-detection sub-routine, active random exploration or passive forgetting mechanism. In Lemma 1, we show a guarantee of RAW-UCB on the favorable event.

Algorithm 1 RAW-UCB

```
Input: $\mathcal{K}$, $\sigma$, $\alpha$
for $t \leftarrow 1, 2, \ldots, K$ do                          ▷ Pull each arm once
    PULL $i_t \leftarrow t$; RECEIVE $o_t$ ; $N_{i_t} \leftarrow 1$
    $\{\widehat{\mu}_{i_t}^h\}_h \leftarrow$ UPDATE($\{\widehat{\mu}_{i_t}^h\}_h$ , $o_t$)          ▷ cf. (1)
end for
for $t \leftarrow K+1, K+2, \ldots$ do
    PULL $i_t \in \arg\max_i \min_{h \le N_i} \widehat{\mu}_i^h + c(h,\delta_t)$   ▷ cf. (3)
    RECEIVE $o_t$ ; $N_{i_t} \leftarrow N_{i_t} + 1$
    $\{\widehat{\mu}_{i_t}^h\}_h \leftarrow$ UPDATE($\{\widehat{\mu}_{i_t}^h\}_h$ , $o_t$)          ▷ cf. (1)
end for
```

Lemma 1. *At round t on favorable event ξ_t^α, if arm i_t is selected, for any $h \le N_{i,t-1}$, the average of its h last pulls cannot deviate significantly from the best available arm at that round, i.e.,*

$$\overline{\mu}_{i_t}^h(t,\pi) \ge \max_{i \in \mathcal{K}} \mu_i(t, N_{i,t-1}) - 2c(h,\delta_t).$$

Seznec et al. (2019) show a slightly worse guarantee about the algorithm FEWA (π_{F}) for the rested rotting bandits. In Appendix C (see Lemma 2), we restate their result using only Assumption 1. FEWA uses the same statistics than RAW-UCB but in a rather complex expanding filtering mechanism which leads to a guarantee of only 4 confidence bounds. Lemma 1 is the only characterization we need for our analysis. Therefore, all our upper bounds will hold for both FEWA and RAW-UCB with their associated constant,

$$C_{\pi_{\mathrm{R}}} \triangleq 2\sqrt{2\alpha} \quad C_{\pi_{\mathrm{F}}} \triangleq 4\sqrt{2\alpha}. \quad (4)$$

Algorithmic complexity FEWA and RAW-UCB have $\mathcal{O}(Kt)$ per round time and space complexity. In Appendix D, we describe EFF-RAW-UCB (π_{ER}) and EFF-FEWA (π_{EF}), two algorithms which reduces the complexities to $\mathcal{O}\left(K\log_m(t)\right)$. It is a refinement of the trick of Seznec et al. (2019) where we add a parameter $m > 1$ to trade-off between complexity and efficiency[3]. For $m = 2$, we prove Lemma 3 and Prop. 11, which are comparable with Lemma 1 and Prop. 2. Therefore, our analysis also holds for these algorithms with,

$$C_{\pi_{\mathrm{ER}}} \triangleq \frac{4\sqrt{\alpha}}{\sqrt{2}-1} \quad C_{\pi_{\mathrm{EF}}} \triangleq \frac{8\sqrt{\alpha}}{\sqrt{2}-1}. \quad (5)$$

[3] When $m < 1 + \frac{1}{T}$, EFF-RAW-UCB behave as RAW-UCB.

The efficient algorithms use less statistics than the original ones. Thus, the probability of the unfavorable event is bounded by $\mathcal{O}\left(t^{1-\alpha}\right)$ (see Prop. 11) which is smaller than $\mathcal{O}\left(t^{2-\alpha}\right)$ in Prop. 2. Hence, our theory holds for a wider range of α for the efficient algorithms.

4 Restless rotting bandits

In this section, the reward decreases independently of the user actions. Hence, we have that $\mu_i(t,n) = \mu_i(t)$.

Variation budget bandits

Setup. Besbes et al. (2014) introduce the limited variation budget bandits, a restless setting where at each round Nature can modify the reward value of any arm but with a limited total variation budget V_T at round T. We combine this assumption with Assumption 1,

Assumption 2. *$\mu_i : \mathbb{N}^\star \to [-V_T, 0]$ are decreasing functions of the time t with V_T a positive constant. Moreover, we have that,*

$$\sum_{t=1}^{T-1} \sup_{i\in\mathcal{K}} \left(\mu_i(t) - \mu_i(t+1)\right) \le V_T\,. \tag{6}$$

Remark 1. *In the rotting scenario, the budget assumption is very similar to the bounded assumption. Indeed, any set of decreasing functions $\mu_i : \mathbb{N}^\star \to [-V, 0]$ satisfies Equation 6 with $V_T = KV$. Reciprocally, any set of functions satisfying Equation 6 with $\mu_i(1) \in [-V_T, 0]$ are bounded in $[-2V_T, 0]$.*

Lower bound. We show that our additional decreasing assumption does not change the minimax rate for budget bandits. This is an adaptation of the proof of Besbes et al. (2014) where we only use rotting function.

Proposition 3. *For any strategy π, there exists a rotting variation budget bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 2 with a budget $V_T \ge \sigma\sqrt{\frac{K}{8T}}$ such that,*

$$\mathbb{E}\left[R_T(\pi)\right] \ge \frac{1}{16\sqrt{2}}\left(\sigma^2 V_T K T^2\right)^{1/3}.$$

Upper bound. `RAW-UCB` matches this lower bound up to poly-logarithmic factors without any knowledge of the horizon T nor the budget V_T.

Theorem 1. *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \ge 4$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \ge 3$ and $m = 2$. For any variation budget bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumptions 2 with variation budget V_T, π suffers an expected regret,*

$$\mathbb{E}\left[R_T(\pi)\right] \le 4\left(C_\pi^2\sigma^2 V_T K T^2 \log T\right)^{1/3} + \widetilde{\mathcal{O}}\left(\left(\sigma V_T^2 K^2 T\right)^{1/3}\right).$$

The remaining terms are of second order when $KV_T \le \mathcal{O}(T)$, which is a necessary condition for the problem to be learnable (see Proposition 3).

Piece-wise stationary bandits.

Setup. In this section, we also consider bounded functions. Hence, they also satisfy Assumption 2 (see Remark 1). However, we further restrained them to be piece-wise stationary,

Assumption 3. *Let V be a positive constant and Υ_T a positive integer. $\mu_i : \mathbb{N}^\star \to [-V, 0]$ are piece-wise stationary non-increasing functions of the time t with at most $\Upsilon_T - 1$ breakpoints.*

Formally, $\sum_{t=1}^{T-1} \mathbb{1}\left(\exists i \in \mathcal{K}, \mu_i(t) \neq \mu_i(t+1)\right) \le \Upsilon_T - 1$. We call $\{t_k\}_{k\le\Upsilon-1}$ the set of breakpoints with $t_0 = 0$, μ_i^k the value of $\mu_i(t)$ for $t \in \{t_k + 1, \ldots, t_{k+1}\}$. We call $i_k^\star \in \arg\max_{i\in\mathcal{K}} \mu_i^k$ (one of) the best arm in batch k and $\Delta_{i,k} \triangleq \mu_{i_k^\star}^k - \mu_i^k$ the gap to the best arm for arm i during batch k. Note that we relax all the assumptions related to the distance between consecutive breakpoints (e.g. Besson and Kaufmann (2019) and their Assumption 4 and 7; Liu et al. (2018) and their Assumption 1 and 2; Cao et al. (2019) and their Assumption 1).

Lower bound. We show that our additional Assumption 1 does not decrease the minimax rate of Garivier and Moulines (2011).

Proposition 4. *For any strategy π, there exists a rotting piece-wise stationary bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 3 with $\Upsilon_T \le \left(\frac{32V^2T}{K\sigma^2}\right)^{1/3}$ such that,*

$$\mathbb{E}\left[R_T(\pi)\right] \ge \frac{\sigma}{32}\sqrt{\Upsilon_T K T}\,.$$

The condition on Υ_T in Proposition 4 follows from Remark 1: if V is too small compared to Υ_T, then we have a budget constraint (with associated lower bound in Proposition 3) rather than a break-point constraint.

Upper bound. `RAW-UCB` matches the lower bound from Proposition 4 up to poly-logarithmic factors without any knowledge of the horizon T nor the number of breakpoints $\Upsilon_T - 1$.

Theorem 2. *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \ge 4$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \ge 3$ and $m = 2$. For any piece-wise stationary bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 3 with $\Upsilon_T - 1$ change-points, π suffers an expected regret,*

$$\mathbb{E}\left[R_T(\pi)\right] \le C_\pi \sigma\sqrt{\log T}\left(\sqrt{\Upsilon_T K T} + \Upsilon_T K\right) + 6KV.$$

Are rotting restless bandits easier? Learning at the minimax rate without knowing Υ_T or V_T was achieved in the non-rotting setup by significantly more complex algorithms. For instance, Auer et al. (2019) use a combination of filtering on the set of potentially good arms, forced exploration planning on identified bad arms, and full restart of the algorithm when a change is detected. This algorithmic complexity has a performance cost, as `AdSwitch` is guaranteed to achieve 56 times the leading term in Theorem 2. Moreover, these algorithms rely on doubling trick when the horizon is unknown, which also has a regret cost compared to intrinsically anytime algorithms (Besson and Kaufmann, 2018).

Yet, Proposition 3 and 4 show that the rotting assumption do not improve the minimax rate for the two considered setups. Interestingly both these lower bounds are matched by (tuned) `Exp3.S` (Auer et al., 2003), an algorithm originally designed for switching best arm in adversarial sequence of rewards. This is comparable to the fixed best arm world: adversarial and stochastic bandits share the same minimax rate which is matched in both setups by `Exp3`. The main interest of the stochastic assumption is to allow for *problem dependent analysis*. For the stochastic stationary bandits, it leads to a stronger $\mathcal{O}(\log(T))$ bounds. In the piece-wise stationary setting, Garivier and Moulines (2011) show that such bounds cannot be achieved without sacrificing the minimax optimality.

Proposition 5 (Theorem 31.2, Lattimore and Szepesvári (2020))**.** *If a policy π performs a regret $R_T(\pi, \mu)$ on a 2-arm stationary instance μ, one can find a piece-wise stationary instance μ' with only two breakpoints such that, for a sufficiently long horizon T, the regret is lower bounded by*

$$\mathbb{E}\left[R_T(\pi, \mu')\right] \geq \frac{T}{22 R_T(\pi, \mu)}.$$

Corollary 1. *Let π a minimax policy on the (non-rotting) piece-wise stationary setups. Then, for a sufficiently large horizon T, there exists a universal constant C such that for all the 2-arm stationary problems μ,*

$$\mathbb{E}\left[R_T(\pi, \mu)\right] \geq C\sqrt{T}.$$

The proof of Proposition 5 is instructive. It builds a problem μ' on which the reward function equals the reward of the stationary problem μ except on a time span τ. During this time span, the best arm of μ keeps its value while the worst arm *increases* to become optimal. The size of τ is chosen inversely proportional to the average pulling rate of the bad arm in μ. Indeed, the lower the pulling rate of the bad arm, the longer the adversary can increase its value in μ' without being noticeable by the learner. Since the pulling rate of the bad arm in μ is proportional to $R_T(\mu)$, we get a lower bound proportional to $\tau \sim \frac{T}{R_T(\mu)}$.

The decreasing Assumption 1 excludes this μ' from the set of possible problems. Theorem 3 shows that not only `RAW-UCB` is able to recover the $\mathcal{O}(\log(T))$ on stationary problems but also recovers the same rate on each batch of a rotting piece-wise stationary problem.

Theorem 3. *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \geq 4$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \geq 3$ and $m = 2$. For any piece-wise stationary bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 3 with $\Upsilon_T - 1$ change-points, π suffers an expected regret*

$$\mathbb{E}\left[R_T(\pi)\right] \leq \sum_{k=0}^{\Upsilon_T - 1} \sum_{i \in \mathcal{K}} \frac{C_\pi^2 \sigma^2 \log T}{\Delta_{i,k}} + \mathcal{O}\left(\sigma \Upsilon_T K \sqrt{\log T}\right).$$

Like in `UCB1` 's analysis, Proposition 2 uses a union-bound with Hoeffding inequality. This technique leads to conservative theoretical tuning of confidence levels and to a suboptimal constant factor $C_\pi^2/2$. One can get the asymptotic optimal tuning for `UCB` on stationary gaussian bandits with a refined analysis which uses a specific concentration result on the deviation of the index (e.g. Lemma 8.2, Lattimore and Szepesvári (2020)). Yet, extending this result to our more complex meta-index and to our several setups is not straightforward and we leave it as future work. Interestingly, the experimental tuning $\alpha = 1.4$ is very close to the asymptotic tuning of `UCB` (see Section 6). It suggests that, besides our union bound considers more events than `UCB` in the theory, we do not have to be significantly more conservative on the confidence levels in practice.

Notice that Mukherjee and Maillard (2019) use a different assumption to get a similar problem-dependent bound. Indeed, they assume that all the arms change at the same time which also excludes μ' from the set of possible problems.

Proofs sketch (full proofs in Appendix E)

Lower bounds. Our proof technique make a strong connection between Proposition 3 and 4. Yet, we adapt existing proofs to the decreasing case (Garivier and Moulines, 2011; Besbes et al., 2014). Hence, we defer the full proof and its sketch to Appendix E.

Upper bounds. We start by separating the regret on the bad events $\overline{\xi_t^\alpha}$ from the good events ξ_t^α. According to Proposition 2, the bad events $\overline{\xi_t^\alpha}$ have low probability for appropriate α. For $\alpha = 4$, they weigh at most $\mathcal{O}(KV)$ in the expected regret. On the good events, we write:

$$R_T(\pi) = \sum_{t=1}^{T} \mu_{i_t^\star}(t) - \overline{\mu}_{i_t}^{h_t}(t, \pi) + \overline{\mu}_{i_t}^{h_t}(t, \pi) - \mu_{i_t}(t)\,. \quad (7)$$

Notice that Lemma 1 can bound the first difference for any h_t. When the reward is piece-wise stationary, we can select h_t such that we include all the pulls of arm i_t from the current stationary batch. If there is none, then it is the first pull of arm i_t in this batch. We handle these $\mathcal{O}(K\Upsilon_T)$ rounds separately (see Lemma 6 in Appendix E). In the other cases, we note that the second difference is null because $\overline{\mu}_{i_t}^{h_t}(t,\pi) = \mu_{i_t}(t) = \mu_i^k$ by the piece-wise stationary assumption. The remaining of the proofs of Theorem 2 and 3 are then very similar to the analysis of Auer et al. (2002) on each stationary batch. Indeed, the two confidence bounds guarantee of Lemma 1 is similar to `UCB1`'s guarantee.

In the variation budget setting, there is no stationary batches. Hence, we cannot choose an h_t which cancels the second difference in Equation 7. Yet, we still decompose the rounds in Υ batches of equal length for the analysis. We choose h_t such that we include all the pulls of arm i_t from the current batch. For the sum of the first differences in Equation 7, there is no difference with the piece-wise stationary case and we can bound

$$\sum_{t=1}^{T} \mu_{i_t^\star}(t) - \overline{\mu}_{i_t}^{h_t}(t,\pi) \leq \widetilde{\mathcal{O}}\left(\sqrt{K\Upsilon T}\right). \tag{8}$$

We call $\Delta_i^k \triangleq \mu_i(t_k) - \mu_i(t_{k+1})$, the total variation of arm i in batch k. The sum of second differences in Equation 7 can be bounded as follows: on each batch of $T\Upsilon^{-1}$ rounds, each second difference is bounded by $\max_{i\in\mathcal{K}} \Delta_i^k$. When we sum over the batches, we get

$$\sum_{t=1}^{T} \overline{\mu}_{i_t}^{h_t}(t,\pi) - \mu_{i_t}(t) \leq \frac{T}{\Upsilon}\sum_{k=0}^{\Upsilon-1} \max_{i\in\mathcal{K}} \Delta_i^k \leq \frac{TV_T}{\Upsilon}. \tag{9}$$

Indeed, in the middle term, we have a maximum on the summed variation of arm i in batch k. On the right-hand side, we have V_T which bounds the sum over the rounds of maximal variation of the arms (see Equation 6). Thus, the right-hand side is larger because the maximum of sums is smaller than the sum of maximums. We can then choose $\Upsilon = \widetilde{\mathcal{O}}\left(T^{1/3}V_T^{2/3}K^{-1/3}\right)$ to minimise the sum of Equation 8 and 9. It leads to the leading term of our Theorem 1. Notice that we still have to handle the first pull of each arm in each batch. If we bound roughly each first pull by V_T, we would get $K\Upsilon V_T \sim \widetilde{\mathcal{O}}\left(V_T^{5/3}\right)$ which would be the leading term for large V_T. Our Lemma 6 is more careful such that it leads to a second order term when $KV_T \leq o(T)$.

5 Rested rotting bandits

Setup We use the rotting setup of Seznec et al. (2019), which extends the one of Levine et al. (2017). This setup is *rested* non-stationary bandits: the change in arm's reward is triggered by the pulls. Hence, we have $\mu_i(t,n) = \mu_i(n)$. Thus, we note that $\overline{\mu}_i^h(t,\pi) = \overline{\mu}_i^h(N_{i,\,t-1}) = \frac{1}{h}\sum_{s=0}^{h-1}\mu_i(N_{i,\,t-1}-s)$. With a slight abuse of notations, we will also use $\widehat{\mu}_i^h(N_{i,\,t-1}) \triangleq \widehat{\mu}_i^h(t,\pi)$[4]. Let

$$L \triangleq \max_{i\in\mathcal{K}} \max_{n\in\{0,\ldots,T-1\}} \mu_i(n) - \mu_i(n-1),$$
$$\text{with } \mu_i(-1) \triangleq \max_{j\in\mathcal{K}} \mu_j(0). \tag{10}$$

Hence, L bounds both the variation of μ_is between two consecutive pulls and the gaps between arms at the first pulls. This is an important quantity for the rested rotting analysis because the minimax rate for the noise-free case is $\mathcal{O}(KL)$ (Heidari et al., 2016).

Theoretical guarantees The analysis of `RAW-UCB` is straightforward from the analysis of `FEWA` due to their similarity. Thus, we recover the problem independent and dependent bounds (see Seznec et al. (2019) for a sketch of the proof, and App. F for a detailed analysis).

Proposition 6 (gap-free bound). *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \geq 5$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \geq 4$ and $m = 2$. For any rotting bandit scenario with means $\{\mu_i\}_i$ satisfying Assumption 1 with bounded decay L and any time horizon T, π suffers an expected regret,*

$$\mathbb{E}\left[R_T(\pi)\right] \leq C_\pi \sigma\sqrt{\log(T)}\left(\sqrt{KT}+K\right) + 6KL.$$

Proposition 7 (gap-dependent bound). *$\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \geq 5$ (or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \geq 4$ and $m = 2$) suffers an expected regret,*

$$\mathbb{E}\left[R_T(\pi)\right] \leq \sum_{i\in\mathcal{K}}\left(\frac{C_\pi^2\sigma^2\log(T)}{\Delta_{i,h_{i,\,T}^+-1}} + C_\pi\sigma\sqrt{\log(T)} + 6L\right)$$

with $h_{i,T}^+ \triangleq \max\left\{h \leq 1 + \frac{C_\pi^2\sigma^2\log T}{\Delta_{i,h-1}^2}\right\}$, and the pseudo-gap

$$\Delta_{i,h} \triangleq \min_{j\in\mathcal{K}} \mu_j\left(N_{j,T}^\star - 1\right) - \overline{\mu}_i^h\left(N_{i,T}^\star + h\right).$$

`RAW-UCB` matches the minimax rate (Prop. 6) up to poly-logarithmic factors. `RAW-UCB` improves over `FEWA`'s problem-dependent guarantee by a factor 4 (Prop. 7). Following Remark 1 of Seznec et al. (2019), one can identify $\Delta_{i,h} = \Delta_i$ in the stationary setting. It gives almost the same guarantee than in Theorem 3 when $\Upsilon_T = 1$ (stationary case). The difference comes from the increased α for the rested case. Indeed, in the rested case, the regret at each round t can be as bad as Lt. Hence, we reduce the probability of the bad event $\overline{\xi_t^\alpha}$ (see Prop. 2). When the reward means are bounded (e.g. for Bernoullis), we can decrease the lower bound on α by one in Propositions 6 and 7.

[4] The average of the observations depends on the realization of the noise ε_t at time t. Yet, these h samples of noise are i.i.d. and thus do not perturb the analysis (see Prop. 2).

6 Real-word data experiment on Yahoo! Front Page

R6A - Yahoo! Front page today module user click log dataset This dataset was used for the Exploration and Exploitation Challenge[5] at ICML 2012 and inspired new algorithms. Among them we mention the work of Tracà and Rudin (2015) who noticed the non-stationary trend and took advantage of it. Since then the dataset continues to be a benchmark[6] for non-stationary bandits (Liu et al., 2018; Cao et al., 2019). It contains the history of clicks on news articles of 45 millions users in the first ten days of May 2009. We use three features in this dataset: *timestamp* (rounded every 5 minutes), *article_id*, and *click*.

A real decaying scenario Every day, between 6pm and 6am EST (12 hours), we notice a decreasing trend in click probability. It suggests that people in the US read less and less news during the evening and night. For every day, we keep all the articles which have been recommended at every timestamp during the 12 hours. For these articles, we use a rolling average window of 30000 in order to estimate the probability of click for each article at each timestamp [7]. We use the real total traffic for each timestamp. We highlight that *we do not enforce any of our assumptions* to create reward functions to be aligned with our setup. In particular, we do not enforce them to be piecewise constant nor to be decreasing. At each round, the learner receives 10 reward samples in order to reduce the cost of computation.

Algorithms and Parameters. We compare RAW-UCB, FEWA, Exp3.S and GLR-UCB. We refer to Appendix G for a discussion about missing algorithms and tuning. Note that our goal is to compare algorithms with the same tuning in the rested and restless benchmark.

Results We display the results for two different days. On day 2, there are several switches of optimal arms with many near-optimal ones: tracking the best arm is an "hard" problem. On day 7, one arm consistently dominates the others by far. Hence, it is an "easy" case where good algorithms should have a logarithmic regret rate. We show the six other days and running time in App. G.2.

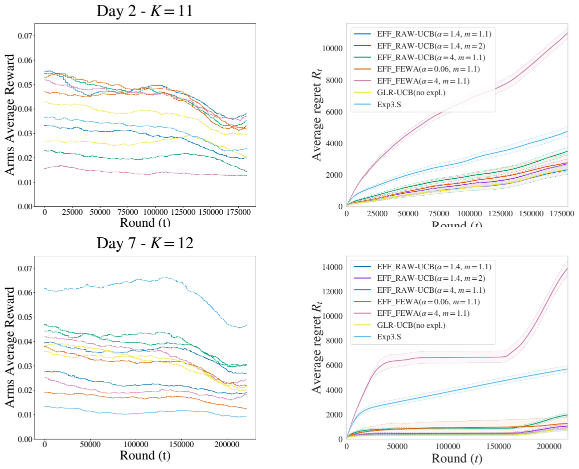


Figure 1: *Left:* rewards from the Yahoo! dataset for two days. *Right:* average regret over 500 runs.

RAW-UCB vs FEWA. The two algorithms compute the same statistics and share most of their analysis. Yet, RAW-UCB consistently outperforms FEWA on the full (rested and restless) benchmark. The difference between the two is even more significant in the restless case. Moreover, RAW-UCB is also simpler to implement and faster to run. Its theoretical tuning $\alpha = 4$ gets reasonable result, while theoretical FEWA is impractical. Finally, its empirical tuning $\alpha_{\mathrm{R}} = 1.4$ is similar to the asymptotic optimal tuning of UCB and shows good performance on both rested and restless problems. By contrast, FEWA with $\alpha_{\mathrm{F}} = 0.06$ shows worse performance with larger deviation on the restless benchmark.

RAW-UCB vs Exp3.S. In Appendix G.1, we show that random exploration of Exp3.S leads to high regret rate in rested rotting bandits. Unsurprisingly, Exp3.S recover more reasonable performance on the restless benchmark, on which it has theoretical guarantees. Yet, it is consistently outperformed by RAW-UCB when we tune the confidence bounds. It is particularly true on easy instance, e.g. on day 7. Indeed, on these cases, we expect logarithmic regret rate for RAW-UCB.

RAW-UCB vs GLR-UCB (no active exploration). GLR-UCB shows good results on the rested benchmark though it is less consistent than RAW-UCB. On the restless benchmark, GLR-UCB shows similar result than RAW-UCB. Yet, we highlight that 1) GLR-UCB needs the knowledge of the horizon to tune its change-detector; 2) we use an efficient version of RAW-UCB which runs ~ 10 times faster than GLR-UCB. In fact, the two algorithms are similar: they are UCB index policies, they recover logarithmic rate on easy restless rotting bandits problems and hence they would both suffer near-linear worst case regret rate in the general restless setting (when active exploration is turned off for GLR-UCB). The main difference is that RAW-UCB scans its history to select its rotting UCB's window, while GLR-UCB scans its history to detect significant changes and restart.

[5] http://explochallenge.inria.fr/

[6] As it allows for offline evaluations as the actions were samples uniformly.

[7] For each timestamp, we average the values given by rolling average. These values are close to each other because the number of click opportunity per article in the same timestamp is small compared to 30000.

Acknowledgements We thank Lilian Besson for hosting and reviewing our numerical experiments code on his bandits package in python (Besson, 2018). The computational experiments were conducted using the Grid'5000 experimental testbed (Balouek et al., 2013).

References

Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of the Conference on Learning Theory (COLT), 2009*, pages 217–226.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2003). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auer, P., Gajane, P., and Ortner, R. (2019). Adaptively Tracking the Best Bandit Arm with an Unknown Number of Distribution Changes. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 138–158, Phoenix, USA. PMLR.

Balouek, D., Amarie, A. C., Charrier, G., Desprez, F., Jeannot, E., Jeanvoine, E., Lèbre, A., Margery, D., Niclausse, N., Nussbaum, L., Richard, O., Perez, C., Quesnel, F., Rohr, C., and Sarzyniec, L. (2013). Adding Virtualization Capabilities to the Grid'5000 Testbed. In *Communications in Computer and Information Science*, volume 367 CCIS, pages 3–20. Springer Verlag.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 199–207. Curran Associates, Inc.

Besson, L. (2018). SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python. Online at: \url{GitHub.com/SMPyBandits/SMPyBandits}.

Besson, L. and Kaufmann, E. (2018). What Doubling Tricks Can and Can't Do for Multi-Armed Bandits.

Besson, L. and Kaufmann, E. (2019). The Generalized Likelihood Ratio Test meets klUCB: an Improved Algorithm for Piece-Wise Non-Stationary Bandits.

Bifet, A. and Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In *Proceedings of the 7th SIAM International Conference on Data Mining*, pages 443–448.

Cao, Y., Wen, Z., Kveton, B., and Xie, Y. (2019). Nearly Optimal Adaptive Procedure with Change Detection for Piecewise-Stationary Bandit. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 418–427. PMLR.

Chapelle, O. and Li, L. (2011). An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 2249–2257.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. (2019). A New Algorithm for Non-stationary Contextual Bandits: Efficient, Optimal and Parameter-free. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 696–726, Phoenix, USA. PMLR.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2019). Learning to Optimize under Non-Stationarity. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1079–1087. PMLR.

Chow, Y. S. and Teicher, H. (1997). *Probability theory : independence, interchangeability, martingales.* Springer.

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

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT), 2011, Espoo, Finland.*, volume 6925 LNAI, pages 174–188. Springer, Berlin, Heidelberg.

Heidari, H., Kearns, M., and Roth, A. (2016). Tight Policy Regret Bounds for Improving and Decaying Bandits. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1562–1570.

Immorlica, N. and Kleinberg, R. (2018). Recharging bandits. In *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, volume 2018-Octob, pages 309–319. IEEE Computer Society.

Komiyama, J. and Qin, T. (2014). Time-Decaying Bandits for Non-stationary Systems. In Liu, T.-Y., Qi, Q., and Ye, Y., editors, *Web and Internet Economics (WINE)*, pages 460–466, Cham. Springer International Publishing.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press UK.

Levine, N., Crammer, K., and Mannor, S. (2017). Rotting Bandits. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 3074–3083.

Liu, F., Lee, J., and Shroff, N. B. (2018). A change-detection based framework for piecewise-stationary multi-armed bandit problem. In McIlraith, S. A. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3651–3658. AAAI Press.

Louëdec, J., Rossi, L., Chevalier, M., Garivier, A., and Mothe, J. (2016). Algorithme de bandit et obsolescence : un modèle pour la recommandation (regular paper). In *Conférence francophone sur l'Apprentissage Automatique, Marseille, 05/07/2016-07/07/2016*, page (en ligne), http://www.lif.univ-mrs.fr. Laboratoire d'Informatique Fondamentale de Marseille.

Mukherjee, S. and Maillard, O.-A. (2019). Distribution-dependent and Time-uniform Bounds for Piecewise i.i.d Bandits.

Pike-Burke, C. and Grunewalder, S. (2019). Recovering Bandits. In H. Wallach and H. Larochelle and A. Beygelzimer and F. d'Alché-Buc and E. Fox and R. Garnett, editor, *Advances in Neural Information Processing Systems 32*, pages 14122—-14131. Curran Associates, Inc.

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted Linear Bandits for Non-Stationary Environments. In Wallach, H., Larochelle, H., Beygelzimer, A., d\textquotesingle Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 12040–12049. Curran Associates, Inc.

Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. (2019). Rotting bandits are no harder than stochastic ones. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research, The 22nd International Conference on Artificial Intelligence and Statistics, 16-18 April 2019.*, volume 89 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR.

Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294.

Tracà, S. and Rudin, C. (2015). Regulating Greed Over Time.

Warlop, R., Lazaric, A., and Mary, J. (2018). Fighting Boredom in Recommender Systems with Linear Reinforcement Learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 1757–1768. Curran Associates, Inc.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. (2016). Tracking the Best Expert in Non-stationary Stochastic Environments. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 3972–3980. Curran Associates, Inc.

Whittle, P. (1980). Multi-Armed Bandits and the Gittins Index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42:143–149.

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

Zimmert, J. and Seldin, Y. (2019). An Optimal Algorithm for Stochastic and Adversarial Bandits. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 467–475. PMLR.

A Outline

The appendix of this paper is organized as follow:

- □ Appendix B is dedicated to the unlearnability of the general decreasing setup.
- □ Appendix C is dedicated to Proposition 2 and Lemma 1. For completion, we also restate a similar Lemma about algorithm FEWA (Seznec et al., 2019) in the rested and restless rotting framework.
- □ Appendix D is dedicated to an efficient version of RAW-UCB.
- □ Appendix E provides the analysis of RAW-UCB for restless rotting bandits.
- □ Appendix F provides the analysis of RAW-UCB for rested rotting bandits.
- □ Appendix G provides all the experiments.

B The general decreasing setup is unlearnable

Proposition 1. *In the no noise setting ($\sigma = 0$), there exists a rotting 2-arms bandits problem (satisfying Assumption 1) with reward value in $[0, 1]$, with one rested arm and one restless arm, and with at most one change-point before T each, such that the greedy oracle strategy π_O suffers a regret*

$$R_T\left(\pi_O\right) \geq \left\lfloor \frac{T}{4} \right\rfloor .$$

Moreover, for any learning strategy π_S, there exists a rotting 2-arms bandits problem (satisfying Assumption 1) with reward value in $[0, 1]$, with one rested arm and one restless arm, and with at most one change-point before T each, such that

$$R_T\left(\pi_S\right) \geq \left\lfloor \frac{T}{8} \right\rfloor .$$

Proof. Let μ^0 and μ^1, two decreasing 2-arms bandits problems such that:

$$\begin{aligned}
&\mu_1^0(t,n) = \mu_1(n) = 1 \text{ if } n < \frac{T}{2} \text{ else } 0\,,\\
&\mu_1^1(t,n) = 1\,,\\
&\mu_2^0(t,n) = \mu_2^1(t,n) = \mu_2(t) = 1/2 \text{ if } t < \frac{T}{2} \text{ else } 0.
\end{aligned}$$

Problem μ^1 only evolves according to time. Hence, the oracle greedy policy π_O is optimal for this problem and collects

$$J_T\left(\pi_O, \mu^1\right) = T. \tag{11}$$

On μ^0, π_O selects arm 1 during $\left\lfloor \frac{T}{2} \right\rfloor$ rounds and then both arms yield 0 reward. Thus, π_O collects

$$J_T\left(\pi_O, \mu^0\right) = \left\lfloor \frac{T}{2} \right\rfloor .$$

However, let π_0 the policy which selects arm 2 for $\left\lfloor \frac{T}{2} \right\rfloor$ rounds and arm 1 afterwards. Thus, π_0 collects

$$J_T\left(\pi_0, \mu^0\right) = (3/2)\left\lfloor \frac{T}{2} \right\rfloor . \tag{12}$$

Hence, we conclude the first part of our proposition,

$$R_T\left(\pi_O, \mu^0\right) = J_T\left(\pi_T^\star, \mu^0\right) - J_T\left(\pi_O, \mu^0\right) \geq J_T\left(\pi_0, \mu^0\right) - J_T\left(\pi_O, \mu^0\right) \geq \left\lfloor \frac{T}{4} \right\rfloor .$$

Now, we consider any learning policy π_S and we call $\mathbb{E}_j[N_{i,t}(\pi_S)]$ the (expected, if the policy is random) number of pulls of arm i at round t by π_S on problem j. Note that the leaner will receive the same rewards for both problems until at least $\left\lfloor \frac{T}{2} \right\rfloor$. Therefore, we have that

$$\forall t \leq \left\lfloor \frac{T}{2} \right\rfloor, \pi(\mathcal{H}_t\left(\mu^0\right)) = \pi(\mathcal{H}_t\left(\mu^1\right)) \implies \mathbb{E}_0\left[N_{2,\left\lfloor \frac{T}{2} \right\rfloor}(\pi_S)\right] = \mathbb{E}_1\left[N_{2,\left\lfloor \frac{T}{2} \right\rfloor}(\pi_S)\right] \triangleq n_2.$$

On problem μ^1, π_S collects a reward of at most,

$$J_T\left(\pi_S,\mu^1\right) = \mathbb{E}_1[N_{1,T}(\pi_S)] + \frac{n_2}{2} = T - \mathbb{E}_1[N_{2,T}(\pi_S)] + \frac{n_2}{2} \leq T - \frac{n_2}{2}, \tag{13}$$

because $n_2 = \mathbb{E}_1\left[N_{2,\lfloor\frac{T}{2}\rfloor}(\pi_S)\right] \leq \mathbb{E}_1[N_{2,T}(\pi_S)]$. Using Equations 11 and 13, we can lower bound the regret of π_S,

$$R_T\left(\pi_S,\mu^1\right) = J_T\left(\pi_O,\mu^1\right) - J_T\left(\pi_S,\mu^1\right) \geq \frac{n_2}{2}.$$

On problem μ^0, π_S collects a reward of at most,

$$J_T\left(\pi_S,\mu^0\right) = \min\left(\mathbb{E}_1[N_{1,T}(\pi_S)], \left\lfloor\frac{T}{2}\right\rfloor\right) + \frac{n_2}{2} \leq \left\lfloor\frac{T}{2}\right\rfloor + \frac{n_2}{2}. \tag{14}$$

Using Equations 12 and 14, we can lower bound the regret of π_S,

$$R_T\left(\pi_S,\mu^0\right) = J_T\left(\pi_O,\mu^0\right) - J_T\left(\pi_S,\mu^0\right) \geq \frac{\lfloor T/2\rfloor - n_2}{2}.$$

Hence, the worst case regret on the two setups is bounded by

$$R_T(\pi_S) \geq \max\left(\frac{n_2}{2}, \frac{\lfloor\frac{T}{2}\rfloor - n_2}{2}\right) \geq \left\lfloor\frac{T}{8}\right\rfloor.$$

□

C Statistical guarantees: Proposition 2 and Lemma 1

Proposition 2. *For any round t and confidence $\delta_t \triangleq 2t^{-\alpha}$, let*

$$\begin{aligned}\xi_t^\alpha \triangleq \Big\{&\forall i \in \mathcal{K},\ \forall n \leq t-1,\ \forall h \leq n,\\ &|\widehat{\mu}_i^h(t,\pi) - \overline{\mu}_i^h(t,\pi)| \leq c(h,\delta_t)\Big\}\end{aligned} \tag{2}$$

be the event under which the estimates at round t are all accurate up to $c(h,\delta_t) \triangleq \sqrt{2\sigma^2\log(2/\delta_t)/h}$. Then, for a policy π which pulls each arms once at the beginning, and for all $t > K$,

$$\mathbb{P}\left[\overline{\xi_t^\alpha}\right] \leq \frac{Kt^2\delta_t}{2} = Kt^{2-\alpha}\,.$$

Proof. We want to upper bound the probability

$$\mathbb{P}\left[\overline{\xi_t^\alpha}\right] = \mathbb{P}\left[\exists i \in K, \exists n \leq t-1, \exists h \leq\ n, |\widehat{\mu}_i^h(t,\pi) - \overline{\mu}_i^h(t,\pi)| > c(h,\delta_t)\right].$$

By Doob's optional skipping (e.g. see Chow and Teicher (1997), Section 5.3) there exists a sequence of random independent variables $(\varepsilon'_l)_{l\in\mathbb{N}}$, σ^2 sub-Gaussian such that

$$\begin{aligned}\mathbb{P}\Big[\exists n \leq t-1, \exists h \leq\ n, &|\widehat{\mu}_i^h(t,\pi) - \overline{\mu}_i^h(t,\pi)| > c(h,\delta_t)\Big]\\ &= \mathbb{P}\left[\exists n \leq t-1, \exists h \leq n,\ |\widehat{\varepsilon}_n^h| > c(h,\delta_t)\right]\\ &\leq \sum_{n=1}^{t-1}\sum_{h=1}^{n}\mathbb{P}\left[|\widehat{\varepsilon}_n^h| > c(h,\delta_t)\right]\\ &\leq \frac{t(t-1)}{2}\cdot\delta_t\,,\end{aligned}$$

where we used the Chernoff inequality in the last line. Thus, a union bound over the arms allows us to conclude that

$$\mathbb{P}\left[\overline{\xi_t^\alpha}\right] \leq \frac{K\delta_t t^2}{2}.$$

□

Lemma 1. *At round t on favorable event ξ_t^α, if arm i_t is selected, for any $h \le N_{i,t-1}$, the average of its h last pulls cannot deviate significantly from the best available arm at that round, i.e.,*

$$\overline{\mu}_{i_t}^{h}(t,\pi) \ge \max_{i\in\mathcal{K}} \mu_i(t, N_{i,t-1}) - 2c(h,\delta_t).$$

Proof. We denote by $i_t^\star \in \arg\max_{i\in\mathcal{K}} \mu_i(t, N_{i,t-1})$, a best available arm at time t and

$$h_{i,t}^{\min} \in \underset{h\le N_{i,t-1}}{\arg\min}\, \widehat{\mu}_i^h(t,\pi) + c(h,\delta_t),$$

a window which minimizes `RAW-UCB` index at time t for arm i. Hence, because the reward functions are non-increasing, we know that

$$\mu_{i_t^\star}(t, N_{i_t^\star,t-1}) \le \overline{\mu}^1_{i_t^\star}(t,\pi) \le \dots \le \overline{\mu}_{i_t^\star}^{h_{i_t^\star,t}^{\min}}(t,\pi)\cdot$$

On the high-probability event ξ_t, we know that the true average of the means cannot deviate significantly from the average of the observed quantity,

$$\overline{\mu}_{i_t^\star}^{h_{i_t^\star,t}^{\min}}(t,\pi) \le \widehat{\mu}_{i_t^\star}^{h_{i_t^\star,t}^{\min}}(t,\pi) + c(h_{i_t^\star,t}^{\min},\delta_t).$$

We know that the selected arm i_t at time t has the largest index, hence,

$$\widehat{\mu}_{i_t^\star}^{h_{i_t^\star,t}^{\min}}(t,\pi) + c(h_{i_t^\star,t}^{\min},\delta_t) \le \widehat{\mu}_{i_t}^{h_{i_t,t}^{\min}}(t,\pi) + c(h_{i_t,t}^{\min},\delta_t).$$

From $h_{i,t}^{\min}$ definition, we know that this quantity is below any upper-confidence bound for any other window h

$$\widehat{\mu}_{i_t}^{h_{i_t,t}^{\min}}(t,\pi) + c(h_{i_t,t}^{\min},\delta_t) \le \widehat{\mu}_{i_t}^{h}(t,\pi) + c(h,\delta_t).$$

Finally, using again the concentration of the average on the ξ_t^α,

$$\widehat{\mu}_{i_t}^{h}(t,\pi) + c(h,\delta_t) \le \overline{\mu}_{i_t}^{h}(t,\pi) + 2c(h,\delta_t).$$

Hence, putting all the equations together, we can write

$$\overline{\mu}_{i_t}^{h}(t,\pi) \ge \max_{i\in\mathcal{K}} \mu_i(t, N_{i,t-1}) - 2c(h,\delta_t).$$

□

For completion, we also restate a similar Lemma about algorithm `FEWA` (Seznec et al., 2019) in the rested and restless rotting framework.

Lemma 2. *For `FEWA` tuned with α, on the favorable event ξ_t^α, if an arm i passes through a filter of window h at round t, i.e., $i \in \mathcal{K}_h$, then the average of its h last pulls satisfies*

$$\overline{\mu}_i^h(t,\pi_{\mathrm{F}}) \ge \max_{i\in\mathcal{K}} \mu_i(t, N_{i,\,t-1}) - 4c(h,\delta_t). \tag{15}$$

Therefore, at round t on favorable event ξ_t^α, if arm i_t is selected by `FEWA` (α), for any $h \le N_{i,\,t-1}$, the average of its h last pulls cannot deviate significantly from the best available arm at that round, i.e.,

$$\overline{\mu}_i^h(t,\pi_{\mathrm{F}}) \ge \max_{i\in\mathcal{K}} \mu_i(t, N_{i,\,t-1}) - 4c(h,\delta_t).$$

Proof. Let $i \in \mathcal{K}_h$ be an arm that passed a filter of window h at round t. First, we use the confidence bound for the estimates and we pay the cost of keeping all the arms up to a distance $2c(h,\delta_t)$ of $\widehat{\mu}^h_{\max,\,t} \triangleq \max_{j\in\mathcal{K}_h} \widehat{\mu}_i^h(t,\pi_{\mathrm{F}})$,

$$\overline{\mu}_i^h(t,\pi_{\mathrm{F}}) \ge \widehat{\mu}_i^h(t,\pi_{\mathrm{F}}) - c(h,\delta_t) \ge \widehat{\mu}^h_{\max,t} - 3c(h,\delta_t) \ge \max_{j\in\mathcal{K}_h} \overline{\mu}_j^h(t,\pi_{\mathrm{F}}) - 4c(h,\delta_t), \tag{16}$$

where in the last inequality, we used that for all $j \in \mathcal{K}_h$,

$$\widehat{\mu}^h_{\max,t} \ge \widehat{\mu}_j^h(t,\pi_{\mathrm{F}}) \ge \overline{\mu}_j^h(t,\pi_{\mathrm{F}}) - c(h,\delta_t).$$

Second, we call $t_{i,t} < t$ the last round at which arm i was selected. Since the means of arms are decaying, we know that

$$
\begin{aligned}
\mu_t^+(\pi_\mathrm{F}) &\triangleq \max_{i\in\mathcal{K}} \mu_i(t, N_{i,\,t-1}) && (17)\\
&\leq \mu_{i_t^\star,\,t_{i,t}} = \overline{\mu}_i^1(t,\pi_\mathrm{F}) && (18)\\
&\leq \max_{j\in\mathcal{K}} \overline{\mu}_j^1(t,\pi_\mathrm{F}) = \max_{j\in\mathcal{K}_1} \overline{\mu}_j^1(t,\pi_\mathrm{F}). && (19)
\end{aligned}
$$

Third, we show that the largest average of the last h' means of arms in $\mathcal{K}_{h'}$ is increasing with h',

$$
\forall h' \leq h,\ \max_{j\in\mathcal{K}_{h'+1}} \overline{\mu}_j^{h'+1}(t,\pi_\mathrm{F}) \geq \max_{j\in\mathcal{K}_{h'}} \overline{\mu}_j^{h'}(t,\pi_\mathrm{F}).
$$

To show the above property, we remark that thanks to our selection rule, the arm that has the largest average of means, always passes the filter. Formally, we show that $\arg\max_{j\in\mathcal{K}_{h'}} \overline{\mu}_j^{h'}(t,\pi_\mathrm{F}) \subseteq \mathcal{K}_{h'+1}$. Let $i_{\max}^{h'} \in \arg\max_{j\in\mathcal{K}_{h'}} \overline{\mu}_j^{h'}(t,\pi_\mathrm{F})$. Then, for such $i_{\max}^{h'}$, we have

$$
\widehat{\mu}_{i_{\max}^{h'}}^{h'}(t,\pi_\mathrm{F}) \geq \overline{\mu}_{i_{\max}^{h'}}^{h'}(t,\pi_\mathrm{F}) - c(h',\delta_t) \geq \overline{\mu}_{\max,t}^{h'} - c(h',\delta_t) \geq \widehat{\mu}_{\max,t}^{h'} - 2c(h',\delta_t),
$$

where the first and the third inequality are due to concentration of the estimates on ξ_t^α, while the second one is due to the definition of $i_{\max}^{h'}$.

Since the arms are decaying, the average of the last $h'+1$ mean values for a given arm is always greater than the average of the last h' mean values and therefore,

$$
\max_{j\in\mathcal{K}_{h'}} \overline{\mu}_j^{h'}(t,\pi_\mathrm{F}) = \overline{\mu}_{i_{\max}^{h'}}^{h'}(t,\pi_\mathrm{F}) \leq \overline{\mu}_{i_{\max}^{h'}}^{h'+1}(t,\pi_\mathrm{F}) \leq \max_{j\in\mathcal{K}_{h'+1}} \overline{\mu}_j^{h'+1}(t,\pi_\mathrm{F}), \tag{20}
$$

because $i_{\max}^{h'} \in \mathcal{K}_{h'+1}$. Gathering Equations 16, 17, and 20 leads to the first claim of the lemma,

$$
\begin{aligned}
\overline{\mu}_i^h(t,\pi_\mathrm{F}) &\overset{(16)}{\geq} \max_{j\in\mathcal{K}_h} \overline{\mu}_j^h(t,\pi_\mathrm{F}) - 4c(h,\delta_t)\\
&\overset{(20)}{\geq} \max_{j\in\mathcal{K}_1} \overline{\mu}_j^1(t,\pi_\mathrm{F}) - 4c(h,\delta_t)\\
&\overset{(17)}{\geq} \mu_t^+(\pi_\mathrm{F}) - 4c(h,\delta_t).
\end{aligned}
$$

To conclude, we remark that if i is pulled at round t, it means that i passes through all the filters from $h=1$ up to $N_{i,\,t-1}$. Therefore, for all $h \leq N_{i,\,t-1}$,

$$
\overline{\mu}_i^h(t,\pi_\mathrm{F}) \geq \mu_t^+(\pi_\mathrm{F}) - 4c(h,\delta_t). \tag{21}
$$

□

D Efficient algorithms

D.1 The numerical cost of adaptive windows

Seznec et al. (2019) highlight that `FEWA` was significantly improving over state-of-the-art algorithms on the rested rotting bandit problem but these improvements are computationally expensive. Indeed, at each round t, we store, update and compare $\mathcal{O}(t)$ statistics. `RAW-UCB` uses the same statistics than `FEWA` (see Prop 2), and thus has the same complexity.

Indeed, the full update of the statistics can be done at a worst case cost of $\mathcal{O}(t)$. Indeed, each statistics $\widehat{\mu}_i^h$ can be refreshed with a $\mathcal{O}(1)$ operation :

$$
\widehat{\mu}_i^{h+1}(n+1) = \frac{h}{h+1}\widehat{\mu}_i^h(n) + \frac{1}{h+1}o_t\,.
$$

The comparison part in both `FEWA` and `RAW-UCB` is also a $\mathcal{O}(t)$ operations. In `FEWA`, we do a scan based on $\widehat{\mu}_i^h$ for all $i \in \mathcal{K}_h$ with increasing h. Hence, the total number of unitary operation is in $\mathcal{O}(t)$ in the worst case, as it scales with the number of statistics. `RAW-UCB` computes one UCB for each of the $\mathcal{O}(t)$ statistics. For each arm, it selects the minimum UCB as index, which can be done with complexity $\mathcal{O}(t)$. Finally, finding the largest index is an $\mathcal{O}(K)$ operations. Therefore, we can conclude,

Proposition 8. `FEWA` and `RAW-UCB` have a $\mathcal{O}(t)$ worst-case complexity per round t in time and memory.

Hence, handling a large number of windows, which is the main strength of these algorithms to achieve a lower regret, is a significant drawback when it comes to design fast algorithms. In the following, we detail and refine the efficient trick of Seznec et al. (2019) by adding a parameter $m \leq 2$ which trades-off between regret and computational performance.

D.2 The efficient update trick

We detail `EFF_UPDATE`, an update scheme to handle efficiently statistics of different windows. A similar yet different approach has appeared independently in the context of streaming mining (Bifet and Gavaldà, 2007). `EFF_UPDATE` is built around two main ideas.

First, at any time t we can avoid using $\left\{\widehat{\mu}_i^h\right\}_h$ for all possible windows h starting from 1 with an increment of 1. In fact, both statistics $\widehat{\mu}_i^h$ and constructed confidence levels $c(h, \delta_t)$ have very close value for successive h as h becomes large :

$$\widehat{\mu}_i^{h+1}(t,\pi) = \widehat{\mu}_i^h(t,\pi) + \mathcal{O}\left(\frac{\sigma + L}{h}\right),$$
$$c(h+1,\delta_t) = c(h,\delta_t) + \mathcal{O}\left(\frac{\sigma}{h^{3/2}}\right).$$

Hence, in both `FEWA` and `RAW-UCB`, we compute a lot of very similar quantities. Instead, we could use fewer statistics which are significantly different : $\left\{\widehat{\mu}_i^h(N_{i,t-1})\right\}_{h \in H_{i,m}}$, where the window h is dispatched on a geometric grid,

$$H_{i,m}(N_{i,t-1}) \triangleq \{h_j \in \{1,\dots,N_{i,t-1}\} \mid h_{j+1} = \lceil m \cdot h_j \rceil \text{ and } h_1 = 1\} \quad \text{with } m > 1.$$

When there is no confusion, we drop the dependency in $N_{i,t-1}$ and use $H_{i,m}$. This modification alone is not enough to reduce both the time and space complexity. Indeed, updating $\widehat{\mu}_i^h$ requires to replace the h-th last sample by the new one o_t. Hence, we need to store all the collected statistics to be able to update all the $\widehat{\mu}_i^h$ for all h with $\mathcal{O}(1)$ complexity. Therefore, in `EFF_UPDATE`, we will use $\mathcal{O}(K\log(t))$ *delayed* statistics that we can update with $\mathcal{O}(K\log(t))$ space and time complexity.

`EFF_UPDATE` (Alg. 2) takes as input the new observation o_t that the learner gets at the N_i-th pull of arm i; the geometric window grid $H_{i,m}$ tuned with an hyperparameter $m > 1$, and for each window h_j in this grid, three different numbers $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$, $p_i^{h_j}$, $n_i^{h_j}$. $\left\{\widehat{\mu}_{i,\texttt{eff}}^{h_j}\right\}_{i,h_j}$ represents the set of *current* statistics of window size h_j that will be used instead of $\left\{\widehat{\mu}_i^h\right\}_{i,h}$ in our efficient algorithms. We also store a pending statistic $p_i^{h_j}$ and a count $n_i^{h_j}$ which are used in the sparse update procedure of $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$. `EFF_UPDATE` outputs an updated set of statistics.

The core of `EFF_UPDATE` is divided in four parts:

1. From Lines 1 to 6, we create new statistics at a logarithmic rate with respect to the growth of N_i;
2. From Lines 7 to 9, we update the statistics of window $h_1 = 1$;
3. From Lines 10 to 13, we update the other pending statistics and count;
4. From Lines 14 to 20, we eventually update $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ and refresh the correspounding pending statistic and count.

The remaining details are quite technical. Thus, we first give the high-level properties that are ensured by the recursive usage of `EFF_UPDATE`. Then, we prove them by going through the algorithm line by line.

Algorithm 2 Eff_Update

```
Input: $o_t$, $H_{i,m} \leftarrow \{h_j < \lceil m \cdot N_i \rceil \mid h_{j+1} = \lceil m \cdot h_j \rceil \text{ with } h_0 = 1\}$, $\left\{\{\widehat{\mu}^{h_j}_{i,\texttt{eff}}, p_i^{h_j}, n_i^{h_j}\}\right\}_{h_j \in H_{i,m}}$
 1: if $N_i = \max(H_{i,m})$ then        ▷ Create a new triplet with window $h_j = \lceil m \cdot N_i \rceil$
 2:     $H_{i,m} \leftarrow H_{i,m} \cup \{\lceil m \cdot N_i \rceil\}$
 3:     $p_i^{\lceil m \cdot N_i \rceil} = p_i^{N_i}$
 4:     $n_i^{\lceil m \cdot N_i \rceil} \leftarrow n_i^{N_i}$
 5:     $\widehat{\mu}^{\lceil m \cdot N_i \rceil}_{i,\texttt{eff}} \leftarrow$ None
 6: end if
 7: $p_i^1 \leftarrow o_t$        ▷ Update the first triplet with $o_t$
 8: $n_i^1 \leftarrow 1$
 9: $\widehat{\mu}^1_{i,\texttt{eff}} \leftarrow o_t$
10: for $h_j \in H_{i,m} \setminus \{1\}$ do        ▷ Update the other pending statistics $p_i^{h_j}$ and $n_i^{h_j}$
11:     $p_i^{h_j} \leftarrow p_i^{h_j} + o_t$
12:     $n_i^{h_j} \leftarrow n_i^{h_j} + 1$
13: end for
14: for $h_j \in$ Sort_Desc$(H_{i,m} \setminus \{1\})$ do
15:     if $n_i^{h_j} = h_j$ then
16:         $\widehat{\mu}^{h_j}_{i,\texttt{eff}} \leftarrow p_i^{h_j}/h_j$        ▷ Replace the current statistic $\widehat{\mu}^{h_j}_{i,\texttt{eff}}$
17:         $p_i^{h_j} = p_i^{h_{j-1}}$        ▷ Refresh the pending statistics
18:         $n_i^{h_j} \leftarrow n_i^{h_{j-1}}$
19:     end if
20: end for
Output: $\left\{\left\{\widehat{\mu}^{h_j}_{i,\texttt{eff}}, p_i^{h_j}, n_i^{h_j}\right\}\right\}_{h_j \in H_{i,m}}$
```

Proposition 9. $\left\{\left\{\widehat{\mu}^{h_j}_{i,\texttt{eff}}, p_i^{h_j}, n_i^{h_j}\right\}\right\}_{h_j \in H_{i,m}}$, *constructed recursively with* `EFF_UPDATE` *with initial value* $\left\{\left\{\widehat{\mu}^1_{i,\texttt{eff}} : \textit{None}, p_i^1 : 0, n_i^1 : 0\right\}\right\}$ *have the following properties :*

1. $\widehat{\mu}^{h_j}_{i,\texttt{eff}}$ *is the average of exactly* h_j *consecutive samples among the* $2h_j - 1$ *last ones.*

2. *The delay between two updates of* $\widehat{\mu}^{h_j}_{i,\texttt{eff}}$ *is in* $\left\{\left\lceil \frac{m-1}{m} h_j \right\rceil, \ldots, h_j - 1\right\}$.

3. *When* $m = 2$, $h_j = 2^j$. *Moreover, for* $j \geq 1$, *the* k*-th update* $\widehat{\mu}^{h_j}_{i,\texttt{eff}}$ *happens at pull* $(k+1) \cdot 2^{j-1}$, i.e. *every* 2^{j-1} *pulls (and at every rounds for* $j = 0$*).*

4. $p_i^{h_j}$ *is the sum of the* $n_i^{h_j}$ *last samples.*

5. $n_i^{h_j} < h_j$ *for* $j \geq 1$. *Also,* $n_i^1 \leq 1$.

6. $\left\{n_i^{h_j}\right\}_{h_j}$ *is an non-decreasing sequence with respect to* h_j *(or* j*).*

Proof. The three last properties are trivially true at the initialization. Thus, we show by induction that they remain true after updates.

Proof of 4. At Lines 3 and 4, we create a new pending statistics and count by initializing them with other statistics and counts. Hence, because of the recursion hypothesis, all the pending statistics $p_i^{h_j}$ (including the created one) contains the sum of the $n_i^{h_j}$ *before last* pulls. At Lines 7 and 8, we update p_i^1 with the last sample and set n_i^1 to 1. At Lines 11 and 12, we add the last sample to $p_i^{h_j}$ (which was containing the before last samples) and increase the count by 1. Hence, at the end of Line 12, all the $p_i^{h_j}$ contain the sum of the last $n_i^{h_j}$ samples. Thus, refreshing $p_i^{h_j}$ and $n_i^{h_j}$ with $p_i^{h_{j-1}}$ and $n_i^{h_{j-1}}$ keeps this property true (Lines 17 and 18).

Proof of 5. For $j = 0$, n_i^1, which is equal to 0 at the initialization, is set at 1 at every update (Line 8). Hence, we have $n_i^{h_0} \leq h_0 = 1$. For $j \geq 1$, $n_i^{\lceil m \cdot N_i \rceil}$ is initialized at Line 4 with the value $n_i^{N_i} < N_i < \lceil m \cdot N_i \rceil$ by the induction hypothesis and because $m > 1$. Then, $n_i^{h_j} < h_j$ $(j \geq 1)$ is increased by one at each update at Line 12. Hence, we now have $n_i^{h_j} \leq h_j$ for all $j \in H_{i,m}$. However, for $j \geq 1$, if $n_i^{h_j} = h_j$ (Line 15), it is replaced by the precedent count $n_i^{h_{j-1}} \leq h_{j-1} < h_j$ (Line 12). Thus, at the end of the update, we do have $n_i^{h_j} < h_j$ for $j \geq 1$.

Proof of 6. At Line 4, we create a new pending count corresponding to the largest h_j and we initialize it with the precedent largest count. At Lines 8 and 12, we set $n_i^1 = 1$ and increase all the other $n_i^{h_j}$ by one. This operation preserves the non-decreasing property of the ordered set. Last, at Line 18, we set few counts $n_i^{h_j}$ to the precedent value $n_i^{h_{j-1}}$- which also preserves the non-decreasing property of the ordered set.

Proof of 1 and 2. Thanks to Property 4, we know that $p_i^{h_j}$ is the sum of the $n_i^{h_j}$ last sample. It is still true at the end of Line 12 (see the proof). Then, at Line 16, and given the condition in Line 15, we set $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ with the average of the last h_j sample. Then, $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ is not updated untill the condition at Line 15 is fulfilled again.

$n_i^{h_j}$ is refreshed with a quantity larger or equal to 1 and smaller or equal to h_{j-1} at Line 18. Then, it is increased by one at each update. we know that $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ will be updated at least every $h_j - 1$, and at most every $h_j - h_{j-1}$ round. Hence, considering the worst possible delay we can conclude : $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ is the average of exactly h_j consecutive samples among the $2h_j - 1$ last ones. Last, considering that $h_{j-1} \leq h_j/m$, we conclude that the minimal delay is larger or equal to $\frac{m-1}{m} h_j$.

Proof of 3. When $m = 2$, it is easy to find by induction that,

$$h_{j+1} = \lceil m \cdot h_j \rceil = 2h_j = 2^{j+1}.$$

For $j = 0$, $\widehat{\mu}_{i,\texttt{eff}}^1$ is updated at every update at Line 9. By induction on $j \geq 1$, $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ is initialized (Line 16) for the first time after $h_j = 2^j = 4 \cdot 2^{j-2}$ pulls. Therefore, it is also an updating pull for $\widehat{\mu}_{i,\texttt{eff}}^{h_{j-1}}$ (by the induction hypothesis) and n_j is set with $n_{j-1} = 2^{j-1}$ at Line 18. Notice that we sort $H_{i,m}$ in the decreasing order at Line 14, hence n_j is updated with n_{j-1} before it is itself updated with n_{j-2}. Hence, $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ is updated again in $h_j - 2^{j-1} = 2^{j-1}$ pulls, *i.e.* after $6 \cdot 2^{j-2}$ pulls of arm i. Again, n_j is set with $n_{j-1} = 2^{j-1}$ (because it is an updating pull for $\widehat{\mu}_{i,\texttt{eff}}^{h_{j-1}}$). By induction, we see that the k-th update happens at pull $(k+1) \cdot 2^{j-1}$, *i.e.* every 2^{j-1} pulls.

□

Remark 2. *At Line 18, we refresh $n_i^{h_j}$ with $n_i^{h_{j-1}}$ which is often larger than 1. Indeed, we could refresh $p_i^{h_j}$ and $n_i^{h_j}$ at 0. Yet, in order to reduce the delay in the update, we use the variable available in the memory which contains the sums of h last sample, with the largest $h < h_j$. According to Properties 4, 5 and 6, this quantity is $p_i^{h_{j-1}}$.*

Notice that we also sort $H_{i,m}$ in the decreasing order at Line 14 to minimize the delay: if there is two consecutive updates of $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ and $\widehat{\mu}_{i,\texttt{eff}}^{h_{j+1}}$ at the same run of `EFF_UPDATE`*, doing a backward loop guarantees to refresh $n_i^{h_{j+1}}$ with a larger value than with a forward loop.*

D.3 `EFF-FEWA` and `EFF-RAW-UCB`

`EFF-FEWA` (π_{EF}) and `EFF-RAW-UCB` (π_{ER}) are the two efficient versions of our initial algorithms. With an hyper-parameter $m > 1$, they use `EFF_UPDATE` instead of `UPDATE` (Lines 3 and 8 in `RAW-UCB`). Therefore, they use $\left\{\widehat{\mu}_{i,\texttt{eff}}^{h_j}\right\}_{i, h_j \in H_{i,m}}$ instead of $\left\{\widehat{\mu}_i^h\right\}_{i, h \leq N_{i,t-1}}$. More precisely, in `RAW-UCB`, we only change the $h \leq N_i$ by $h_j \in H_{i,m}$ and $\widehat{\mu}_i^h$ by $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ in the index computation at Line 7. We can perform similar changes to adapt `FEWA` in `EFF-FEWA`.

Proposition 10. `EFF-FEWA` and `EFF-RAW-UCB` tuned with hyper-parameter m have a $\mathcal{O}(K \log_m(t))$ worst-case time and space complexity at round t.

Proof. The total number of statistics for each arm i at round t is bounded by $\mathcal{O}\left(\log_m(t)\right)$. Indeed,

$$t \geq N_{i,t-1} \geq h_j \geq m^{j-1} \implies j \leq 1 + \log_m(t)\,.$$

Moreover, in `EFF_UPDATE` we use 3 numbers for each $\left\{\widehat{\mu}_{i,\texttt{eff}}^{h_j}\right\}_j$. Hence, the space complexity scales with

$$\sum_{i\in\mathcal{K}} |H_{i,m}| = \sum_{i\in\mathcal{K}} \mathcal{O}\left(\log_m(t)\right) = \mathcal{O}\left(K\log_m(t)\right).$$

The time complexity of `EFF_UPDATE` scales with the number of statistics in arm i_t, *i.e.* at most $\mathcal{O}\left(\log_m(t)\right)$. The indexes computation of `EFF-RAW-UCB` find the minimum of K sets with cardinality $\mathcal{O}\left(\log_m(t)\right)$, while finding the maximum among these indexes is a $\mathcal{O}(K)$ operation. Thus, the worst-case time complexity is $\mathcal{O}\left(K\log_m(t)\right)$. `EFF-FEWA` uses at most $\mathcal{O}\left(\log_m(t)\right)$ times the procedure `FILTER` whose inner complexity scales with $|\mathcal{K}_h| \leq K$. Therefore, in the worst case, the time complexity of `EFF-FEWA` at round t is bounded by $\mathcal{O}\left(K\log_m(t)\right)$. □

D.4 Analysis

The analysis of `RAW-UCB` and `FEWA` only uses Proposition 2 and Lemma 4. We will derive analogous results for `EFF-RAW-UCB` and `EFF-FEWA`. The upper-bounds will directly follow with no additional effort.

A favorable event for efficiently updated adaptive windows

Proposition 11. *For any round t and confidence $\delta_t \triangleq 2t^{-\alpha}$, let*

$$\xi_{t,m}^{\alpha} \triangleq \left\{\forall i\in\mathcal{K},\ \forall n\leq t-1,\ \forall h_j \in H_{i,m}(n), |\widehat{\mu}_{i,\textit{eff}}^{h_j}(t,\pi) - \overline{\mu}_{i,\textit{eff}}^{h_j}(t,\pi)| \leq c(h_j,\delta_t)\right\}$$

be the event under which the estimates at round t are all accurate up to $c(h,\delta_t) \triangleq \sqrt{2\sigma^2\log(2/\delta_t)/h}$. Then, for a policy π which pulls each arms once at the beginning, and for all $t > K$,

$$\mathbb{P}\left[\overline{\xi_{t,2}^{\alpha}}\right] \leq 3Kt\delta_t = 6Kt^{1-\alpha}\,.$$

Remark 3. *The probability of the unfavorable event $\overline{\xi_{t,2}^{\alpha}}$ scales with $\mathcal{O}\left(t^{1-\alpha}\right)$ compared to $\mathcal{O}\left(t^{2-\alpha}\right)$ for $\overline{\xi_t^{\alpha}}$ because the efficient algorithms construct less statistics. It means that our theory will hold for a wider range of α. Yet, this benefits is only theoretical. The union bound in Proposition 2 is not tight because the different statistics share the same data: the confidence bounds are not independent at all. In practice, it leads to conservative tuning of the confidence bounds and one can decrease α to get better performance.*

Proof. As in Proposition 2, we have to count the number of statistics that are required to hold in the confidence region. Calling $u_j(t)$ the number of update of statistics $\widehat{\mu}_{i,\texttt{eff}}^{h_j}$ after t pulls, we have

$$\begin{aligned}
\mathbb{P}\left[\overline{\xi_{t,2}^{\alpha}}\right] &\leq \sum_{i\in\mathcal{K}} \sum_{j=0}^{\lfloor\log_2(t)\rfloor-1} u_j(t)\delta_t \\
&\leq \sum_{i\in\mathcal{K}} \left(t-1+\sum_{j=1}^{\lfloor\log_2(t)\rfloor-1} \frac{t-1}{2^{j-1}}\right)\delta_t \\
&\leq 3Kt\delta_t
\end{aligned}$$

In the second inequality, we use Property 3 in Proposition 9: statistics $\widehat{\mu}_{i,\texttt{eff}}^{h_j}(n)$ is only updated every 2^{j-1} pulls for $j \geq 1$ (and every pull for $j = 0$). □

Lemma 3. *At round t on favorable event $\xi_{t,2}^{\alpha}$, if arm i_t is selected by $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $m = 2$, for any $h \leq N_{i,t-1}$, the average of its h last pulls cannot deviate significantly from the best available arm at that round, i.e.,*

$$\overline{\mu}_{i_t}^{h}(t-1,\pi) \geq \max_{i\in\mathcal{K}} \mu_i(t, N_{i,t-1}) - \frac{C_\pi}{\sqrt{2\alpha}}c(h,\delta_t) \quad \textit{with} \begin{cases} C_{\pi_{\mathrm{ER}}} = \frac{4\sqrt{\alpha}}{\sqrt{2}-1} \\ C_{\pi_{\mathrm{EF}}} = \frac{8\sqrt{\alpha}}{\sqrt{2}-1} \end{cases}.$$

Proof. Like for Lemma 2 (see its proof), our proof is done in a more general rotting framework that can be used in the next chapter. We denote by $\overline{\mu}_i^{hh'}(t-1,\pi)$ and $\widehat{\mu}_i^{hh'}(t-1,\pi)$ the true mean and empirical average associated to the $h'-h$ samples between the h-th last one (included) and the h'-th last one (excluded). Let $j_h \in \mathbb{N}^\star$ such that : $2^{j_h}-1 \leq h < 2^{j_h+1}$.

$$\overline{\mu}_{i_t}^{h}(t-1,\pi) \geq \overline{\mu}_{i_t}^{2^{j_h}-1}(t-1,\pi) = \sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} \overline{\mu}_{i_t}^{2^j 2^{j+1}}(t-1,\pi). \tag{22}$$

The inequality follows because the reward is decreasing and $h \geq 2^{j_h}-1$. Then, we decompose the average in a weighted sum of averages of geometrically expanding windows. Since the reward is decreasing we have that,

$$\forall k \leq 2^j, \quad \overline{\mu}_{i_t}^{2^j 2^{j+1}}(t-1,\pi) \geq \overline{\mu}_{i_t}^{k:k+2^j}(t-1,\pi).$$

$\widehat{\mu}_{i_t,\texttt{eff}}^{h_j}$ contains 2^j samples among the $2^{j+1}-1$ last ones (see Proposition 9). Setting $k \leq 2^j$ to the current delay of the statistics $\widehat{\mu}_{i_t,\texttt{eff}}^{h_j}$ (see Point 2 in Proposition 9), we can write,

$$\overline{\mu}_{i_t}^{2^j 2^{j+1}}(t-1,\pi) \geq \overline{\mu}_{i_t}^{k:k+2^j}(t-1,\pi) = \overline{\mu}_{i_t,\texttt{eff}}^{h_j} \geq \widehat{\mu}_{i_t,\texttt{eff}}^{h_j} - c(2^j,\delta_t), \tag{23}$$

where we use that we are on $\xi_{t,\,2}^{\alpha}$ at the last line. Therefore, gathering Equations 22 and 23,

$$\overline{\mu}_{i_t}^{h}(t-1,\pi) \geq \sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} \left(\widehat{\mu}_{i_t,\texttt{eff}}^{h_j} - c(2^j,\delta_t)\right). \tag{24}$$

Now, we will use the mechanics of the two algorithms. On the first hand, for `EFF-RAW-UCB`, we make the index appear in the inequality,

$$\begin{aligned}
\overline{\mu}_{i_t}^{h}(t-1,\pi_{\mathrm{ER}}) &\geq \sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} \left(\widehat{\mu}_{i_t,\texttt{eff}}^{h_j} - c(2^j,\delta_t)\right) & (25)\\
&= \sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} \left(\widehat{\mu}_{i_t,\texttt{eff}}^{h_j} + c(2^j,\delta_t) - 2c(2^j,\delta_t)\right) & (26)\\
&\geq \min_{j\in H_{i,\,2}} \left(\widehat{\mu}_{i_t,\texttt{eff}}^{h_j} + c(2^j,\delta_t)\right) - 2\sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} c(2^j,\delta_t). & (27)
\end{aligned}$$

Then, we can relate the left part of the sum to the best current value $\mu_{i_t^\star}(t, N_{i_t^\star,\,t-1})$,

$$\min_{j\in H_{i,\,2}} \left(\widehat{\mu}_{i_t,\texttt{eff}}^{h_j} + c(2^j,\delta_t)\right) \geq \min_{j\in H_{i_t^\star,2}} \left(\widehat{\mu}_{i_t^\star,\,\texttt{eff}}^{h_j} + c(2^j,\delta_t)\right) \geq \overline{\mu}_{i_t^\star,\,\texttt{eff}}^{h_{\min}} \geq \mu_{i_t^\star}(t, N_{i_t^\star,\,t-1}). \tag{28}$$

where $h_{\min} \in \arg\min_{h_j \in H_{i,\,2}} \left(\widehat{\mu}_{i,\texttt{eff}}^{h_j} + c(h_j,\delta_t)\right)$.The first inequality follows because `EFF-RAW-UCB` selects the arm with the largest index. In particular, the index of i_t is larger or equal to the index of $i_t^\star \in \arg\max_{i\in\mathcal{K}} \mu_i(t, N_{i_t^\star,\,t})$. The second inequality holds on $\xi_{t,\,2}^{\alpha}$. The third inequality uses the decreasing of the reward. Putting Equations 27 and 28, we get,

$$\overline{\mu}_{i_t}^{h}(t-1,\pi_{\mathrm{ER}}) \geq \mu_{i_t^\star}(t, N_{i_t^\star,\,t-1}) - 2\sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} c(2^j,\delta_t). \tag{29}$$

On the other hand, for `EFF-FEWA`, we know that the selected arm passes any filter of window $2^j \in H_{i,\,2}$. Therefore, with $i_{\max} \in \arg\max_{i\in\mathcal{K}_{h_j}} \overline{\mu}_{i,\texttt{eff}}^{h_j}$, we can write,

$$\begin{aligned}
\widehat{\mu}_{i_t,\texttt{eff}}^{h_j} &\geq \max_{i\in\mathcal{K}_{h_j}} \widehat{\mu}_{i,\texttt{eff}}^{h_j} - 2c\left(h_j,\delta_t\right) & \text{Filtering rule} \quad (30)\\
&\geq \widehat{\mu}_{i_{\max},\,\texttt{eff}}^{h_j} - 2c\left(h_j,\delta_t\right) & i_{\max} \in \mathcal{K}_{h_j} \quad (31)\\
&\geq \overline{\mu}_{i_{\max},\,\texttt{eff}}^{h_j} - 3c(h_j,\delta_t) & \text{on } \xi_{t,\,2}^{\alpha} \quad (32)\\
&= \max_{i\in\mathcal{K}_{h_j}} \overline{\mu}_{i,\texttt{eff}}^{h_j} - 3c\left(h_j,\delta_t\right). & (33)
\end{aligned}$$

We relate $\overline{\mu}^{h_j}_{i,\mathtt{eff}}$ to the largest available value at round t,

$$\max_{i\in\mathcal{K}_{h_j}} \overline{\mu}^{h_j}_{i,\mathtt{eff}} \geq \max_{i\in\mathcal{K}_1} \overline{\mu}^{1}_{i,\mathtt{eff}} = \max_{i\in\mathcal{K}} \overline{\mu}^{1}_{i,\mathtt{eff}} \geq \overline{\mu}^{1}_{i_t^\star,\mathtt{eff}} \geq \mu_{i_t^\star}(t, N_{i_t^\star,t-1}). \tag{34}$$

The last inequality follows from the decreasing of the reward and the before last from the definition of the maximum operator. The first one uses a similar argument than in Lemma 2 : $\max_{i\in\mathcal{K}_{h_j}} \overline{\mu}^{h_j}_{i,\mathtt{eff}}$ increases with h_j. Indeed, on $\xi^{\alpha}_{t,2}$,

$$i_j \triangleq \arg\max_{i\in\mathcal{K}_{h_j}} \overline{\mu}^{h_j}_{i,\mathtt{eff}} \in \mathcal{K}_{h_{j+1}},$$

because it cannot be at more than two confidence bounds from the best empirical value during the filter h_j. Thus, we get,

$$\max_{i\in\mathcal{K}_{h_j}} \overline{\mu}^{h_j}_{i,\mathtt{eff}} = \overline{\mu}^{h_j}_{i_j,\mathtt{eff}} \leq \overline{\mu}^{h_{j+1}}_{i_j,\mathtt{eff}} \leq \max_{i\in\mathcal{K}_{h_{j+1}}} \overline{\mu}^{h_{j+1}}_{i,\mathtt{eff}}.$$

The first inequality follows because $\overline{\mu}^{h_{j+1}}_{i_j,\mathtt{eff}}$ contains reward sample which are either in $\overline{\mu}^{h_j}_{i_j,\mathtt{eff}}$ or are older than the ones in $\overline{\mu}^{h_j}_{i_j,\mathtt{eff}}$. Indeed, when $m=2$, $\widehat{\mu}^{h_{j+1}}_{i,\mathtt{eff}}$ is updated synchronously with $\widehat{\mu}^{h_j}_{i,\mathtt{eff}}$ (see Property 3 in Proposition 9). Hence, at each update of $\widehat{\mu}^{h_{j+1}}_{i,\mathtt{eff}}$, it contains all the samples of $\widehat{\mu}^{h_j}_{i,\mathtt{eff}}$ and the 2^j precedent ones. Thus, because the reward is decreasing, we have $\overline{\mu}^{h_{j+1}}_{i_j,\mathtt{eff}} \geq \overline{\mu}^{h_j}_{i_j,\mathtt{eff}}$. The second inequality uses that $i_j \in \mathcal{K}_{h_{j+1}}$. Gathering Equations 24, 33 and 34, we get

$$\overline{\mu}^{h}_{i_t}(t-1,\pi_{\mathrm{EF}}) \geq \mu_{i_t^\star}(t, N_{i_t^\star,t-1}) - 4\sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} c(2^j,\delta_t). \tag{35}$$

With few lines of algebra, we reduce the sum,

$$\begin{aligned}
\sum_{j=0}^{j_h-1} \frac{2^j}{2^{j_h}-1} c(2^j,\delta_t) &= \sum_{j=0}^{j_h-1} \frac{\sqrt{2}^j}{2^{j_h}-1} c(1,\delta_t) && c(2^j,\delta_t) = \frac{c(1,\delta_t)}{\sqrt{2^j}} \\
&= \frac{\sqrt{2}^{j_h}-1}{\left(\sqrt{2}-1\right)\left(2^{j_h}-1\right)} c(1,\delta_t) && \sum_{n=0}^{N} q^n = \frac{q^{N+1}-1}{q-1} \\
&= \frac{1}{\left(\sqrt{2}-1\right)\left(\sqrt{2}^{j_h}+1\right)} c(1,\delta_t) && a^2-1=(a-1)(a+1) \\
&\leq \frac{\sqrt{2}}{\left(\sqrt{2}-1\right)\sqrt{2^{j_h+1}}} c(1,\delta_t) && \sqrt{2^{j_h}}+1 \geq \frac{\sqrt{2^{j_h+1}}}{\sqrt{2}} \\
&\leq \frac{\sqrt{2}}{\left(\sqrt{2}-1\right)\sqrt{h}} c(1,\delta_t) && h \leq 2^{j_h+1} \\
&= \frac{\sqrt{2}}{\sqrt{2}-1} c(h,\delta_t). && \frac{c(1,\delta_t)}{\sqrt{h}} = c(h,\delta_t)
\end{aligned}$$

Plugging this last equation in Equations 29 and 35 leads to the final result,

$$\overline{\mu}^{h}_{i_t}(t,\pi) \geq \max_{i\in\mathcal{K}} \mu_i(t, N_{i,t-1}) - \frac{C_\pi}{\sqrt{2\alpha}} c(h,\delta_t) \quad \text{with} \quad \begin{cases} C_{\pi_{\mathrm{ER}}} = \frac{4\sqrt{\alpha}}{\sqrt{2}-1} \\ C_{\pi_{\mathrm{EF}}} = \frac{8\sqrt{\alpha}}{\sqrt{2}-1} \end{cases}.$$

□

Remark 4. ***Can we adapt our theory for*** $m \neq 2$***?*** For `EFF-FEWA`, we used at Equation 34 that $\widehat{\mu}^{h_j}_{i,\mathtt{eff}}$ is synchronously updated with the precedent statistics which is a specific characteristic for $m=2$. For `EFF-RAW-UCB`, the proof could work using a grid $\{2h_j,\ldots,2h_{j+1}-1\}$ to decompose the means (at Eq. 22). Yet, the computation is much messier, mainly because of the ceil operator in $h_{j+1} = \lceil m \cdot h_j \rceil$. The constant ratio compared to `RAW-UCB`'s guarantee one could get with this technique would be no better than $\sqrt{m}\frac{m-1}{\sqrt{m}-1} = \sqrt{m}\left(\sqrt{m}+1\right)$. When $m \to 1$, this constant does not go to one: it is disappointing because we know that `EFF-RAW-UCB` is equivalent to `RAW-UCB` for $m \leq 1 + \frac{1}{T}$.

Finally, we give a synthetic claim of Lemmas 1, 2 and 3.

Lemma 4. *At round t on favorable event ξ_t^α (respectively, $\xi_{t,2}^\alpha$), if arm i_t is selected by $\pi \in \{\pi_\mathrm{F}, \pi_\mathrm{R}\}$ (respectively, $\pi \in \{\pi_\mathrm{EF}, \pi_\mathrm{ER}\}$ tuned with $m = 2$), for any $h \le N_{i,t-1}$, the average of its h last pulls cannot deviate significantly from the best available arm at that round, i.e.,*

$$\overline{\mu}_{i_t}^h(t,\pi) \ge \max_{i \in \mathcal{K}} \mu_i(t) - \frac{C_\pi}{\sqrt{2\alpha}} c(h, \delta_t) \quad \textit{with} \begin{cases} C_{\pi_\mathrm{R}} = 2\sqrt{2\alpha} \textit{ and } C_{\pi_\mathrm{ER}} = \frac{4\sqrt{\alpha}}{\sqrt{2}-1} \\ C_{\pi_\mathrm{F}} = 4\sqrt{2\alpha} \textit{ and } C_{\pi_\mathrm{EF}} = \frac{8\sqrt{\alpha}}{\sqrt{2}-1} \end{cases}.$$

E Analysis for the restless setting

Lower bounds

The two lower bounds follow the same analysis. We build a set of rotting piece-wise stationary problems with an evenly spaced set of $\Upsilon - 1$ breakpoints. The adversary can choose the distance between arms $\Delta = \frac{1}{4}\sqrt{\frac{\sigma^2 K \Upsilon}{2T}}$ at the maximum such that the best arm is barely identifiable between two breakpoints. Hence, at each break-point, each arm's value decreases by Δ or 2Δ. Even if the set of breakpoints would be known, the learner does not know which arm is the best on each stationary part. Hence, in the worst case, she suffers at least the sum of the minimax regret of Υ stationary bandits problems with horizon $\frac{T}{\Upsilon}$, *i.e.* $\mathcal{O}\left(\sqrt{K\Upsilon T}\right)$. In the piece-wise stationary setting, we can simply identify $\Upsilon = \Upsilon_T$. In the variation budget setting, the adversary has a constraint over $\Upsilon\Delta = \frac{1}{4}\sqrt{\frac{\sigma^2 K \Upsilon^3}{2T}} = \mathcal{O}(V_T)$. Hence, when the budget is limited, the adversary can choose up to $\Upsilon = \mathcal{O}\left(T^{1/3}\right)$ breakpoints such that the sub-optimal arms are "sufficiently" far from the best one (*i.e* at Δ). This dependence on T leads to the increased regret rate of $\mathcal{O}\left(T^{2/3}\right)$.

Lemma 5. *Let $\Upsilon \in \{1, \dots, T\}$ and $\left\{\tau_k \triangleq \left\lceil \frac{T}{\Upsilon} \right\rceil \text{ if } k \le T \bmod \Upsilon \text{ else } \left\lfloor \frac{T}{\Upsilon} \right\rfloor\right\}_{k \le \Upsilon}$. We call $t_k = \sum_{k'=1}^{k} \tau_{k'}$ and $t_0 = 0$. Consider a family of piece-wise stationary bandits indexed by a vector $i^\star \in (\{0\} \cup \mathcal{K})^\Upsilon$ as follows: arm i is a Gaussian distribution $\mathcal{N}(\mu_i(t), \sigma)$ such that*

$$\forall k \in \{0, \dots, \Upsilon - 1\}, \ \forall t \in \{t_{k-1} + 1, \dots, t_k\}, \ \mu_i(t) = \begin{cases} -k\Delta \text{ if } i = i_k^\star \\ -(k+1)\Delta \text{ else.} \end{cases}$$

We denote by $\mathbb{E}_{i^\star}$ the expectation under the problem indexed by $i^\star$. Then, if $\Delta = \frac{1}{4}\sqrt{\frac{\sigma^2 K \Upsilon}{2T}}$, for any policy π :

$$\exists i^\star \in (\{0\} \cup \mathcal{K})^\Upsilon, \ \mathbb{E}_{i^\star}[R_T(\pi)] \ge \frac{\sqrt{\sigma^2 K T \Upsilon}}{32}.$$

Proof. Note that when $i_k^\star = 0$ then all the arms share the same means. We also define the vector $i_{-k}^\star$ equals to $i^\star$ with the coordinate k empty and for $i \in \mathcal{K}$ the vector $(i_{-k}^\star, i)$ as the vector where we fill the empty coordinate with i. We fix a policy π and we will lower bound its average regret on the bandits problem indexed by $i^\star \in \mathcal{K}^\Upsilon$

$$\begin{aligned} \frac{1}{K^\Upsilon} \sum_{i^\star \in \mathcal{K}^\Upsilon} \mathbb{E}_{i^\star}[R_T(\pi)] &= \frac{1}{K^\Upsilon} \sum_{i^\star \in \mathcal{K}^\Upsilon} \sum_{k=1}^{\Upsilon} \Delta \mathbb{E}_{i^\star}[\tau_k - N_{i_k^\star}^k] \\ &= \Delta \left(T - \frac{1}{K^\Upsilon} \sum_{i^\star \in \mathcal{K}^\Upsilon} \sum_{k=1}^{\Upsilon} \mathbb{E}_{i^\star}[N_{i_k^\star}^k] \right), \end{aligned}$$

where N_i^k is the number of pulls of arm i during epoch k. Thus we need to upper bound the following quantity

$$\frac{1}{K^\Upsilon} \sum_{i^\star \in \mathcal{K}^\Upsilon} \sum_{k=1}^{\Upsilon} \mathbb{E}_{i^\star}[N_{i_k^\star}^k] = \sum_{k=1}^{\Upsilon} \frac{1}{K^{\Upsilon-1}} \sum_{i_{-k}^\star \in \mathcal{K}^{\Upsilon-1}} \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{(i_{-k}^\star, i)}[N_i^k].$$

Using the contraction of the entropy for the bounded random variable N_i^k/τ_k then the Pinsker inequality (see Garivier et al., 2019) we get

$$2\left(\frac{1}{\tau_k K} \sum_{i=1}^{K} \mathbb{E}_{(i_{-k}^\star, i)}[N_i^k] - \frac{1}{\tau_k K} \sum_{i=1}^{K} \mathbb{E}_{(i_{-k}^\star, 0)}[N_i^k] \right)^2 \le \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{(i_{-k}^\star, 0)}[N_i^k] \frac{\Delta^2}{2\sigma^2},$$

since problems $(i^\star_{-k}, i)$ and $(i^\star_{-k}, 0)$ differ only by a gap Δ on the arm i during epoch k. Thanks to the fact that $\sum_i N_i^k \leq \tau_k$ we get

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{(i^\star_{-k},i)}[N_i^k] \leq \frac{\tau_k}{K} + \frac{\Delta}{2\sigma\sqrt{K}}\tau_k^{3/2}\,.$$

Putting all together we have for $K \geq 2$

$$\frac{1}{K^\Upsilon}\sum_{i^\star\in\mathcal{K}^\Upsilon}\mathbb{E}_{i^\star}[R_T(\pi)] \geq \left(\frac{T}{2} - \sum_{k=1}^{\Upsilon}\frac{\tau_k^{3/2}\Delta}{2\sigma\sqrt{K}}\right)\Delta\,.$$

We have $\tau_k = \left\lfloor\frac{T}{\Upsilon}\right\rfloor$ or $\tau_k = \left\lceil\frac{T}{\Upsilon}\right\rceil$ such that $\sum_{k=1}^{\Upsilon}\tau_k = T$. Hence, we have that $\tau_k \leq 2T/\Upsilon$ which leads to

$$\frac{1}{K^\Upsilon}\sum_{i^\star\in\mathcal{K}^\Upsilon}\mathbb{E}_{i^\star}[R_T(\pi)] \geq \left(\frac{1}{2}T - \frac{\sqrt{2}T^{3/2}\Delta}{\sigma\sqrt{K\Upsilon}}\right)\Delta\,.$$

Choosing $\Delta = \frac{1}{4}\sqrt{\frac{\sigma^2 K\Upsilon}{2T}}$, we get

$$\frac{1}{K^\Upsilon}\sum_{i^\star\in\mathcal{K}^\Upsilon}\mathbb{E}_{i^\star}[R_T(\pi)] \geq \frac{1}{4}\sqrt{\frac{\sigma^2 K\Upsilon}{2T}}\left(\frac{1}{4}T\right) \geq \frac{\sqrt{\sigma^2 KT\Upsilon}}{32}.$$

We can conclude by noticing that the average expected regret across the problem set is lesser or equal to the maximum across the same problem set. □

Proposition 4. *For any strategy π, there exists a rotting piece-wise stationary bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 3 with $\Upsilon_T \leq \left(\frac{32V^2T}{K\sigma^2}\right)^{1/3}$ such that,*

$$\mathbb{E}\left[R_T(\pi)\right] \geq \frac{\sigma}{32}\sqrt{\Upsilon_T KT}\,.$$

Proof. This result directly follows from Lemma 5 by choosing $\Upsilon = \Upsilon_T$. Indeed, the set of problems $\left\{i^\star \in (\{0\}\cup\mathcal{K})^{\Upsilon_T}\right\}$ satisfy Assumption 3 as soon as $\Upsilon_T\Delta \leq V$, *i.e.* $\Upsilon_T \leq \left(\frac{32V^2T}{K\sigma^2}\right)^{1/3}$. □

Proposition 3. *For any strategy π, there exists a rotting variation budget bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 2 with a budget $V_T \geq \sigma\sqrt{\frac{K}{8T}}$ such that,*

$$\mathbb{E}\left[R_T(\pi)\right] \geq \frac{1}{16\sqrt{2}}\left(\sigma^2 V_T KT^2\right)^{1/3}.$$

Proof. We want to use Lemma 5 but we need to make the set of problems $\left\{i^\star \in (\{0\}\cup\mathcal{K})^{\Upsilon_T}\right\}$ comply with Assumption 2. First, the function are bounded by $-V_T$. Hence, we need :

$$\Upsilon\Delta \leq V_T. \tag{36}$$

Second the total variation is bounded according to Equation 6. When t is not a break-point, the variation is null. At each break-point, the maximal variation across the arm is 2Δ. For $\Upsilon - 1$ break-point, we have that

$$2\Delta\left(\Upsilon - 1\right) \leq V_T. \tag{37}$$

Since $2\Delta\left(\Upsilon-1\right) \leq \frac{\sigma}{2}\sqrt{\frac{K}{2T}}\Upsilon^{3/2}$, we choose

$$\Upsilon = \min\left(\max\left(\left\lfloor 2\left(\frac{V_T^2 T}{K\sigma^2}\right)^{1/3}\right\rfloor, 1\right), T\right). \tag{38}$$

By construction, 38 satisfies 37. Moreover, when $\Upsilon > 1$, 37 is more restrictive than 36. For $\Upsilon = 1$, we simply assume $\Delta \leq V_T$, *i.e.* $V_T \geq \sigma\sqrt{\frac{K}{8T}}$.

Plugging 38 in Lemma 5 allows us to conclude

$$\mathbb{E}\left[R_T(\pi)\right] \geq \frac{1}{16\sqrt{2}} V_T^{1/3}\sigma^{2/3}K^{1/3}T^{2/3}.$$

□

Upper bounds

Lemma 6 (Bound on unfavorable events. Decomposition in unspecified batches. Bound on the first pull of each arm in each batch)**.** *Let an integer $\Upsilon \in \{1, \ldots, T\}$.*
Let $\mu_i : \mathbb{N}^\star \to [0, -V]$, the K decreasing reward functions.
Let $\{t_k \in \{1, \ldots, T\} \mid t_k > t_{k-1}\}_{k\in\{1,\ldots,\Upsilon-1\}}$ a set of $\Upsilon - 1$ distinct rounds delimiting Υ batches. We set $t_0 = 0$ and $t_\Upsilon = T$.
We call $h_i^k \triangleq \sum_{t=t_k+1}^{t_{k+1}} \mathbb{1}\left(i_t = i\right)$ the number of pulls of arm i in batch k and $t_i^k(h)$ the time at which arm i is pulled for the h-th time since $t_k + 1$. We also call $\mathcal{K}_k \triangleq \left\{i \in \mathcal{K} | h_i^k \geq 1\right\}$ the set of pulled arms in batch k.

Then, $\pi \in \{\pi_{\mathrm{R}}, \pi_{\mathrm{F}}\}$ run with $\alpha \geq 4$, or $\pi \in \{\pi_{\mathrm{ER}}, \pi_{\mathrm{EF}}\}$ run with $m = 2$ and $\alpha \geq 3$, suffers an expected regret of

$$\begin{aligned}\mathbb{E}\left[R_T(\pi)\right] \leq & \,\mathbb{E}\left[\sum_{k=0}^{\Upsilon-1}\sum_{i\in\mathcal{K}_k}\sum_{t=t_k+1}^{t_{k+1}}\sum_{h=2}^{h_i^k}\mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right)\Big(\mu_\star(t) - \mu_i(t)\Big)\right] \\ & + C_\pi \sigma \Upsilon K\sqrt{\log T} + 6KV.\end{aligned}$$

Proof. We start by separating the favorable events from the unfavorable events:

$$R_T(\pi) = \underbrace{\sum_{t=1}^{T}\mathbb{1}\left(\xi_t^\alpha\right)\left(\mu_\star(t) - \mu_{i_t}(t)\right)}_{R_T(\pi|\xi_t^\alpha)} + \underbrace{\sum_{t=1}^{T}\mathbb{1}(\overline{\xi_t^\alpha})\left(\mu_\star(t) - \mu_{i_t}(t)\right)}_{R_T(\pi|\overline{\xi_t^\alpha})}, \tag{39}$$

with $\mu_\star(t) \triangleq \max_{i\in\mathcal{K}} \mu_i(t)$. For $\alpha \geq 4$, we can bound the cost of the unfavorable events thanks to Proposition 2,

$$\mathbb{E}\left[R_T(\pi|\overline{\xi_t^\alpha})\right] \leq \sum_{t=1}^{T}\mathbb{P}\left[\overline{\xi_t^\alpha}\right]V \leq \sum_{t=1}^{T}\frac{KV}{t^2} = \frac{KV\pi^2}{6} \leq 2KV. \tag{40}$$

On the favorable events, given any ordered set of $\Upsilon - 1$ breakpoints $\{t_k\}$, we divide the horizon in Υ batches $\{t_k + 1, \ldots, t_{k+1}\}_{k\leq\Upsilon-1}$,

$$R_T(\pi|\xi_t^\alpha) \leq \sum_{k=0}^{\Upsilon-1}\sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(\xi_t^\alpha\right)\left(\mu_\star(t) - \mu_{i_t}(t)\right).$$

We define h_i^k the number of pulls of arm i in batch k, *i.e.* $h_i^k = \sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(i_t = i\right)$. We use $t_i^k(h)$ to designate the time at which arm i is pulled for the h-th time since t_k.

$$R_T(\pi|\xi_t^\alpha) \leq \sum_{k=0}^{\Upsilon-1}\sum_{t=t_k+1}^{t_{k+1}}\sum_{i\in\mathcal{K}_k}\sum_{h=1}^{h_i^k}\mathbb{1}\left(t_i^k(h) = t \wedge \xi_t^\alpha\right)\Big(\mu_\star(t) - \mu_i(t)\Big).$$

We split the regret on the first pulls of each batch

$$\begin{aligned}R_T(\pi|\xi_t^\alpha) =& \underbrace{\sum_{k=0}^{\Upsilon-1}\sum_{t=t_k+1}^{t_{k+1}}\sum_{i\in\mathcal{K}_k}\mathbb{1}\left(t=t_i^k(1)\wedge\xi_t^\alpha\right)\Big(\mu_\star(t)-\mu_i(t))\Big)}_{FP} \\ &+\underbrace{\sum_{k=0}^{\Upsilon-1}\sum_{t=t_k+1}^{t_{k+1}}\sum_{i\in\mathcal{K}_k}\sum_{h=2}^{h_i^k}\mathbb{1}\left(t=t_i^k(h)\wedge\xi_t^\alpha\right)\Big(\mu_\star(t)-\mu_i(t)\Big)}_{OP}.\end{aligned} \tag{41}$$

Analysis of the first pulls. We call k_i^1, the index of the batch at which arm i is pulled for the first time. We call $\mathcal{K}_k^2 \triangleq \left\{i\in\mathcal{K}_k | k > k_i^1\right\}$, the set of arms pulled at least once during batch k and at least once in a batch before k. We split the regret due to the very first pull each arm from the other first pulls in each batch,

$$FP = \sum_{k=0}^{\Upsilon-1}\sum_{i\in\mathcal{K}_k}\sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(t=t_i^k(1)\wedge\xi_t^\alpha\right)\Big(\mu_\star(t)-\mu_i(t)\Big) \tag{42}$$

$$\le \sum_{i\in\mathcal{K}}\Big(0-\mu_i(t_i^{k_i^1}(1))\Big) + \sum_{k=1}^{\Upsilon-1}\sum_{i\in\mathcal{K}_k^2}\sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(t=t_i^k(1)\wedge\xi_t^\alpha\right)\Big(\mu_\star(t)-\mu_i(t)\Big) \tag{43}$$

$$= \sum_{i\in\mathcal{K}}\Big(0-\mu_i(t_i^{k_i^1}(1))\Big) \tag{44}$$

$$+\sum_{k=1}^{\Upsilon-1}\sum_{i\in\mathcal{K}_k^2}\sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(t=t_i^k(1)\wedge\xi_t^\alpha\right)\Big(\mu_\star(t)-\overline{\mu}_i^1(t,\pi)+\overline{\mu}_i^1(t,\pi)-\mu_i(t)\Big). \tag{45}$$

The inequality is justified because $\mu_i(t)\le 0$ for all t. In the last equation, we simply introduce $\overline{\mu}_i^1(t,\pi)$, the last pulled sample of arm i, which is well defined after the first pull of each arm. According to Lemma 4, the first difference is bounded on the high-probability event ξ_t^α,

$$\sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(t=t_i^k(1)\wedge\xi_t^\alpha\right)\left(\mu_\star(t)-\overline{\mu}_i^1(t,\pi)\right) \le \frac{C_\pi}{\sqrt{2\alpha}}c(1,2T^{-\alpha}) = C_\pi\sigma\sqrt{\log T}. \tag{46}$$

We will show that we can telescope the second sum. First, we notice that we can collapse the sum on t using $\mathbb{1}\left(t=t_i^k(1)\right)$. Moreover, ξ_t^α will not be needed: hence we can drop $\mathbb{1}\left(\xi_t^\alpha\right)\le 1$.

$$\sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(t=t_i^k(1)\wedge\xi_t^\alpha\right)\left(\overline{\mu}_i^1(t,\pi)-\mu_i(t)\right) \le \overline{\mu}_i^1(t_i^k(1),\pi)-\mu_i(t_i^k(1)). \tag{47}$$

For a given batch k on which arm i is pulled, the precedent reward sample has a mean $\overline{\mu}_i^h\left(t_i^k\left(1\right),\pi\right)$. This sample is the last pull of the last batch k' before k on which arm i is pulled. Hence, its mean is smaller than the mean of the first pull on this same batch k' because the reward is decreasing. Hence, the sum can telescope

$$\sum_{i\in\mathcal{K}}\Big(0-\mu_i(t_i^{k_i^1}(1))\Big) + \sum_{k=1}^{\Upsilon-1}\sum_{i\in\mathcal{K}_k^2}\sum_{t=t_k+1}^{t_{k+1}}\mathbb{1}\left(t=t_i^k(1)\wedge\xi_t^\alpha\right)\left(\overline{\mu}_i^1(t,\pi)-\mu_i(t)\right) \tag{48}$$

$$\le \sum_{i\in\mathcal{K}}\left\{0-\mu_i(t_i^{k_i^1}(1)) + \sum_{k=k_i^1+1}^{\Upsilon-1}\mathbb{1}\left(h_i^k\ge 1\right)\left(\overline{\mu}_i^1(t_i^k(1),\pi)-\mu_i(t_i^k(1))\right)\right\} \tag{49}$$

$$\le \sum_{i\in\mathcal{K}}\Big(0-\mu_i(T)\Big) \le KV\,. \tag{50}$$

The first inequality uses the definition of $\mathcal{K}_k^2$ along with Equation 47. The second inequality follows from the telescoping argument presented above. The third inequality uses that $\mu_i(T) \geq -V$. Gathering Equation 46 and 48, we can bound the term FP (defined in Equation 41)

$$FP \leq KV + \sum_{k=1}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k^2} C_\pi \sigma \sqrt{\log T} \leq KV + C_\pi \sigma \Upsilon K \sqrt{\log T}. \tag{51}$$

Conclusion. From Equation 39, we can bound the expected regret on the unfavorable events thanks to Equation 40. On the favorable events, we can split the rounds in batches on which we isolate the first pull of each arm on each batch thanks to Equation 41. Finally, we bound the regret due to these first pulls thanks to Equation 51, and for $\alpha \geq 4$,

$$\begin{aligned}\mathbb{E}\left[R_T(\pi)\right] \leq & \mathbb{E}\left[\sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right)\left(\mu_\star(t) - \mu_i(t)\right)\right] \\ & + C_\pi \sigma \Upsilon K \sqrt{\log T} + 3KV.\end{aligned}$$

For the efficient algorithms, we can use the same proof with $\xi_{t,2}^\alpha$ and get for $\alpha \geq 3$,

$$\begin{aligned}\mathbb{E}\left[R_T(\pi)\right] \leq & \mathbb{E}\left[\sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right)\left(\mu_\star(t) - \mu_i(t)\right)\right] \\ & + C_\pi \sigma \Upsilon K \sqrt{\log T} + 6KV.\end{aligned}$$

□

Lemma 7 (Analysis of the second pulls in each batch under the favorable events.). *Let $\Delta_i^k \triangleq \mu_i(t_k+1) - \mu_i(t_{k+1})$, the decrement of arm i in batch k. For any arm i and any consecutive rounds $\{t_k+1, \ldots, t_{k+1}\}$ such that i is pulled $h_i^k \geq 1$ times, the regret due to the pulls after the first one can be bounded under the favorable events,*

$$\sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right)\left(\mu_\star(t) - \mu_i(t)\right) \leq \left(h_i^k - 1\right)\Delta_i^k + \sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi_{t_i^k(h)}^\alpha\right)\left(\mu_\star(t_i^k(h)) - \overline{\mu}_i^{h-1}(t_i^k(h), \pi)\right).$$

Proof. We call $\Delta_i(t, t') \triangleq \mu_i(t) - \mu_i(t')$ the variation of arm i between times t and t'. As a short notation, we refer to $\Delta_i^k \triangleq \Delta_i(t_k+1, t_{k+1})$ for the variation of arm i in batch k.

$$\forall h \leq h_i^k, \quad \mu_i(t_i^k(h)) \geq \mu_i(t_{k+1}) = \mu_i(t_k+1) - \Delta_i^k \geq \overline{\mu}_i^{h-1}(t_i^k(h), \pi) - \Delta_i^k\,. \tag{52}$$

The two inequalities are justified by the rewards decay. Indeed, any pull in batch k has a higher reward than the value of arm i at the end of the batch t_{k+1}. Moreover, the value at the beginning of the batch is higher that any average of h value in this batch. The middle equality follows from the definition of Δ_i^k.

Then, we plug Equation 52 in the left hand side of our claim,

$$\begin{aligned}&\sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right)\left(\mu_\star(t) - \mu_i(t)\right) \\ &\qquad = \sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi_{t_i^k(h)}^\alpha\right)\left(\mu_\star(t_i^k(h)) - \mu_i(t_i^k(h))\right) \\ &\qquad \leq \sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi_{t_i^k(h)}^\alpha\right)\left(\mu_\star(t_i^k(h)) - \overline{\mu}_i^{h-1}(t_i^k(h), \pi) + \Delta_i^k\right) \\ &\qquad \leq \left(h_i^k - 1\right)\Delta_i^k + \sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi_{t_i^k(h)}^\alpha\right)\left(\mu_\star(t_i^k(h)) - \overline{\mu}_i^{h-1}(t_i^k(h), \pi)\right).\end{aligned}$$

The last inequality is justified by $\mathbb{1}\left(\xi^{\alpha}_{t_i^k(h)}\right) \leq 1$. □

Variation budget rotting bandits.

Theorem 1. *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \geq 4$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \geq 3$ and $m = 2$. For any variation budget bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumptions 2 with variation budget V_T, π suffers an expected regret,*

$$\mathbb{E}\left[R_T(\pi)\right] \leq 4\left(C_\pi^2 \sigma^2 V_T K T^2 \log T\right)^{1/3} + \widetilde{\mathcal{O}}\left(\left(\sigma V_T^2 K^2 T\right)^{1/3}\right).$$

Proof. Let $\Upsilon \in \{1, \dots, T\}$ a number of evenly spaced batches that we will specify later. We define the length of these batches $\left\{\tau_k \triangleq \left\lceil \frac{T}{\Upsilon} \right\rceil \text{ if } k \leq T \bmod \Upsilon \text{ else } \left\lfloor \frac{T}{\Upsilon} \right\rfloor\right\}_{k \leq \Upsilon}$. Note that $\sum_{k=1}^{\Upsilon} \tau_k = T$. Let $t_k = \sum_{k'=0}^{k} \tau_{k'}$ the last round of each batch and $t_0 = 0$. On each of these batches, we apply Lemma 7 for the set of arms which have been pulled in this batch,

$$\sum_{k=0}^{\Upsilon_T - 1} \sum_{i \in \mathcal{K}_k} \sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_t^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right) \Big(\mu_\star(t) - \mu_i(t)\Big) \leq \sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \left(h_i^k - 1\right) \Delta_i^k$$
$$+ \sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi^{\alpha}_{t_i^k(h)}\right) \left(\mu_\star(t_i^k(h)) - \overline{\mu}_i^{h-1}(t_i^k(h), \pi)\right). \quad (53)$$

The first sums can be handled using Assumption 2 and the evenly spaced property of τ_k,

$$\sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}} \left(h_i^k - 1\right) \Delta_i^k \leq \sum_{k=0}^{\Upsilon-1} \max_{j \in \mathcal{K}} \Delta_j^k \sum_{i \in \mathcal{K}} \left(h_i^k - 1\right) = \sum_{k=0}^{\Upsilon-1} \max_{j \in \mathcal{K}} \Delta_j^k \left(\tau_k - K\right) \leq \frac{T}{\Upsilon} \sum_{k=0}^{\Upsilon-1} \max_{j \in \mathcal{K}} \Delta_j^k. \quad (54)$$

The first inequality is justified by definition of the maximum. The second equality states that the total number of pulls in batch k is τ_k. The third inequality uses that $\tau_k - K \leq \left\lceil \frac{T}{\Upsilon} \right\rceil - K \leq \left\lceil \frac{T}{\Upsilon} \right\rceil - K \leq \frac{T}{\Upsilon}$. Now, we need to relate $\max_{j \in \mathcal{K}} \Delta_j^k$ and V_T,

$$\sum_{k=0}^{\Upsilon-1} \max_{j \in \mathcal{K}} \Delta_j^k = \sum_{k=0}^{\Upsilon-1} \max_{j \in \mathcal{K}} \sum_{t=t_k+1}^{t_{k+1}-1} \Delta_j(t, t+1) \leq \sum_{k=0}^{\Upsilon-1} \sum_{t=t_k+1}^{t_{k+1}-1} \max_{j \in \mathcal{K}} \Delta_j(t, t+1) \leq \sum_{t=1}^{T} \max_{j \in \mathcal{K}} \Delta_j(t, t+1) \leq V_T. \quad (55)$$

The first inequality is justified because the maximum of a sum is smaller than the sum of the maximums. In the second inequality, we add positive terms which are the maximum of the decay among the arms at the boundary between batches. The last inequality is justified by Assumption 2. Therefore, we can bound the first sums using Equation 54 and 55,

$$\sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}} \left(h_i^k - 1\right) \Delta_i^k \leq \frac{V_T T}{\Upsilon}. \quad (56)$$

The second sums can be bounded using Lemma 4 on the high probability event $\xi^{\alpha}_{t_i^k(h)}$ and Jensen's inequality,

$$\sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi^{\alpha}_{t_i^k(h)}\right) \left(\mu_\star(t_i^k(h)) - \overline{\mu}_i^{h-1}(t_i^k(h), \pi)\right) \leq \sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{h=2}^{h_i^k} \frac{C_\pi c\left(h-1, 2T^{-\alpha}\right)}{\sqrt{2\alpha}} \quad (57)$$

$$= \sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{h=2}^{h_i^k} C_\pi \sigma \sqrt{\frac{\log T}{h-1}} \quad (58)$$

$$\leq \sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} 2 C_\pi \sigma \sqrt{h_i^k \log T} \quad (59)$$

$$\leq 2 C_\pi \sigma \sqrt{\Upsilon K T \log T}. \quad (60)$$

We remark that the bound in Eq. 56 is decreasing with Υ and the bound in Eq. 60 is increasing with Υ. We will choose Υ in order to minimize the sum of these two bounds (which will be our leading term). Therefore, we set,

$$\Upsilon \triangleq \left\lceil \left(\frac{V_T^2 T}{C_\pi^2 \sigma^2 K \log T} \right)^{1/3} \right\rceil . \tag{61}$$

We have that $\Upsilon \leq T$ when $V_T \leq C_\pi \sigma T \sqrt{K \log T}$. Moreover, we will use that $\Upsilon \leq 2 \left(\frac{V_T^2 T}{C_\pi^2 \sigma^2 K \log T} \right)^{1/3}$ which is true when $V_T \geq \sqrt{\frac{C_\pi^2 \sigma^2 K \log T}{8T}}$.

Finally, we use Lemma 6 where we replace the inner sums thanks to Equations 53, 56 and 60. Then, we plug Υ set in 61 and conclude,

$$\begin{aligned}
\mathbb{E}\left[R_T\left(\pi\right)\right] &\leq \frac{V_T T}{\Upsilon} + 2 C_\pi \sigma \sqrt{\Upsilon K T \log T} + C_\pi \sigma \Upsilon K \sqrt{\log T} + 6 V_T K \\
&\leq 4 \left(C_\pi^2 \sigma^2 V_T K T^2 \log T \right)^{1/3} + 2 \Big(C_\pi \sigma V_T^2 K^2 T \sqrt{\log T} \Big)^{1/3} + 6 V_T K .
\end{aligned}$$

When $V_T \leq \sqrt{\frac{C_\pi^2 \sigma^2 K \log T}{8T}}$, the regret of any policy can be bounded ,

$$\begin{aligned}
\mathbb{E}\left[R_T(\pi)\right] \leq T V_T &= V_T^{1/3} T^{2/3} V_T^{2/3} T^{1/3} \\
&\leq V_T^{1/3} T^{2/3} \left(\frac{C_\pi^2 \sigma^2 K \log T}{8T} \right)^{1/3} T^{1/3} \\
&= \frac{1}{2} \left(C_\pi^2 \sigma^2 V_T K T^2 \log T \right)^{1/3} \\
&\leq 4 \left(C_\pi^2 \sigma^2 V_T K T^2 \log T \right)^{1/3} .
\end{aligned}$$

For completion, we also consider $V_T \geq C_\pi \sigma T \sqrt{K \log T}$. Yet, notice that in that case the leading term is $\mathcal{O}\left(K V_T\right)$. We start back from Lemma 6,

$$\begin{aligned}
\mathbb{E}\left[R_T(\pi)\right] \leq & \mathbb{E}\left[\sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha \right) \Big(\mu_\star(t) - \mu_i(t) \Big) \right] \\
& + C_\pi \sigma \Upsilon K \sqrt{\log T} + 6 K V_T .
\end{aligned}$$

In fact, this result can be slightly improved at no cost,

$$\begin{aligned}
\mathbb{E}\left[R_T(\pi)\right] \leq & \mathbb{E}\left[\sum_{k=0}^{\Upsilon-1} \sum_{i \in \mathcal{K}_k} \sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha \right) \Big(\mu_\star(t) - \mu_i(t) \Big) \right] \\
& + C_\pi \sigma \min\left(\Upsilon K, T\right) \sqrt{\log T} + 6 K V_T ,
\end{aligned}$$

because there are at most $\min\left(\Upsilon K, T\right)$ first pulls (see the proof of Lemma 6). Now, we choose $\Upsilon = T$. Hence, there is no second pulls and we have,

$$\mathbb{E}\left[R_T(\pi)\right] \leq C_\pi \sigma T \sqrt{\log T} + 6 K V_T ,$$

Now we use that $C_\pi \sigma T \sqrt{\log T} \leq \frac{V_T}{\sqrt{K}} \leq K V_T$,

$$\begin{aligned}
\mathbb{E}\left[R_T(\pi)\right] &\leq \Big(C_\pi \sigma T \sqrt{\log T} \Big)^{2/3} \Big(C_\pi \sigma T \sqrt{\log T} \Big)^{1/3} + 6 K V_T \\
&\leq \left(C_\pi^2 \sigma^2 V_T K T^2 \log T \right)^{1/3} + 6 K V_T \\
&\leq 4 \left(C_\pi^2 \sigma^2 V_T K T^2 \log T \right)^{1/3} + 2 \Big(C_\pi \sigma V_T^2 K^2 T \sqrt{\log T} \Big)^{1/3} + 6 K V_T .
\end{aligned}$$

□

Piece-wise stationary rotting bandits.

Let $\{t_k\}_{\{k \leq \Upsilon_T\}}$ be the set of breakpoints with $t_0 = 0$ and $t_{\Upsilon_T} = T$. For all $t \in \{t_k+1, \dots, t_{k+1}\}$, $\mu_i(t) = \mu_i^k$. We denote $i_k^\star \in \arg\max_{i \in \mathcal{K}} \mu_i^k$ (one of) the best arm(s) in batch k, and $\mu_\star^k \triangleq \max_{i \in \mathcal{K}} \mu_i^k$, the corresponding best value. We also call $\Delta_{i,k} \triangleq \mu_\star^k - \mu_i^k$ the gap between arm i and optimal arm in batch k.

Lemma 8. *For an arm i and a stationary batch k, we call $h_{i,\xi}^k \triangleq \max\left(h \leq h_i^k | \xi^\alpha_{t_i^k(h)}\right)$ the last pull of arm i in batch k under the favorable events (possibly 0). If $h_{i,\xi}^k \geq 1$, the regret due to the second pulls on the favorable events is bounded by,*

$$\sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right) \left(\mu_\star(t) - \mu_i(t)\right) \leq \left(h_{i,\xi}^k - 1\right) \Delta_{i,k} \leq C_\pi \sigma \sqrt{\left(h_{i,\xi}^k - 1\right) \log T}.$$

Proof. We apply Lemma 7 on each stationary batch. Hence, $\Delta_i^k = 0$ and we can write,

$$\sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right) \left(\mu_\star(t) - \mu_i(t)\right) \leq \sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi^\alpha_{t_i^k(h)}\right) \left(\mu_\star(t_i^k(h)) - \overline{\mu}_i^{h-1}(t_i^k(h), \pi)\right).$$

We notice that $\mu_\star(t_i^k(h)) = \mu_\star^{(k)}$. We call $h_{i,\xi}^k \triangleq \max\left(h \leq h_i^k \; | \xi^\alpha_{t_i^k(h)}\right)$. Hence,

$$\begin{aligned}
\sum_{h=2}^{h_i^k} \mathbb{1}\left(\xi^\alpha_{t_i^k(h)}\right) \left(\mu_\star(t_i^k(h)) - \overline{\mu}_i^{h-1}(t_i^k(h), \pi)\right) &= \sum_{h=2}^{h_{i,\xi}^k} \mathbb{1}\left(\xi^\alpha_{t_i^k(h)}\right) \left(\mu_\star^k - \overline{\mu}_i^{h-1}(t_i^k(h), \pi)\right) \\
&\leq \sum_{h=2}^{h_{i,\xi}^k} \mu_\star^k - \overline{\mu}_i^{h-1}(t_i^k(h), \pi) \\
&= \sum_{h=2}^{h_{i,\xi}^k} \mu_\star^k - \mu_i^k \\
&= \left(h_{i,\xi}^k - 1\right) \Delta_{i,k} \,.
\end{aligned}$$

The first equality follows from $\forall h > h_{i,\xi}^k$, $\mathbb{1}\left(\xi^\alpha_{t_i^k(h)}\right) = 0$ by definition of $h_{i,\xi}^k$. The first inequality follows by dropping $\mathbb{1}\left(\xi^\alpha_{t_i^k(h)}\right) \leq 1$. The second equality uses that the function is stationary in batch k : $\forall h \leq h_{i,\xi}^k, \overline{\mu}_i^{h-1}(t_i^k(h), \pi) = \mu_i^k$. The last equality follows by definition of $\Delta_{i,k}$ (which does not depend on the summand index h).

Then, we apply Lemma 4 at time $t_i^k\left(h_{i,\xi}^k\right)$. By definition of $h_{i,\xi}^k$, $\mathbb{1}\left(\xi^\alpha_{t_i^k(h_{i,\xi}^k)}\right) = 1$.

$$\left(h_{i,\xi}^k - 1\right) \Delta_{i,k} \leq \frac{C_\pi}{\sqrt{2\alpha}} \left(h_{i,\xi}^k - 1\right) c(h_{i,\xi}^k - 1, 2T^{-\alpha}) = C_\pi \sigma \sqrt{\left(h_{i,\xi}^k - 1\right) \log T}.$$

□

Theorem 2. *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \geq 4$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \geq 3$ and $m = 2$. For any piece-wise stationary bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 3 with $\Upsilon_T - 1$ change-points, π suffers an expected regret,*

$$\mathbb{E}\left[R_T(\pi)\right] \leq C_\pi \sigma \sqrt{\log T} \left(\sqrt{\Upsilon_T K T} + \Upsilon_T K\right) + 6KV.$$

Proof. We apply Lemma 8,

$$\sum_{k=0}^{\Upsilon_T - 1} \sum_{i \in \mathcal{K}_k} \sum_{t=t_k+1}^{t_{k+1}} \sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right) \left(\mu_\star(t) - \mu_i(t)\right) \leq \sum_{k=0}^{\Upsilon_T - 1} \sum_{i \in \mathcal{K}_k} C_\pi \sigma \sqrt{h_{i,\xi}^k \log T}.$$

We notice that $\sum_{k=0}^{\Upsilon_T-1}\sum_{i\in\mathcal{K}_k} h_{i,\xi}^k \leq T$. Hence, thanks to Jensen's inequality,

$$\sum_{k=0}^{\Upsilon_T-1}\sum_{i\in\mathcal{K}_k} C_\pi\sigma\sqrt{h_{i,\xi}^k \log T} \leq C_\pi\sigma\sqrt{\Upsilon_T K T\log T}.$$

We use Lemma 6 with the last equation and conclude,

$$\mathbb{E}\left[R_T(\pi)\right] \leq C_\pi\sigma\sqrt{\log T}\left(\sqrt{\Upsilon_T K T} + \Upsilon_T K\right) + 6KV.$$

□

Theorem 3. *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \geq 4$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \geq 3$ and $m = 2$. For any piece-wise stationary bandit scenario with means $\{\mu_i(t)\}_{i,t}$ satisfying Assumption 3 with $\Upsilon_T - 1$ change-points, π suffers an expected regret*

$$\mathbb{E}\left[R_T(\pi)\right] \leq \sum_{k=0}^{\Upsilon_T-1}\sum_{i\in\mathcal{K}} \frac{C_\pi^2\sigma^2\log T}{\Delta_{i,k}} + \mathcal{O}\left(\sigma\Upsilon_T K\sqrt{\log T}\right).$$

Proof. Let $\mathcal{K}_k \triangleq \{i \in \mathcal{K} | \Delta_{i,k} > 0\}$, the set of sub-optimal arms in batch k. We apply Lemma 8 to bound the number of wrong pull (under the favorable events) of arm $i \in \mathcal{K}_k$ during batch k,

$$\Delta_{i,k}\left(h_{i,\xi}^k - 1\right) \leq C_\pi\sigma\sqrt{\left(h_{i,\xi}^k - 1\right)\log T} \implies h_{i,\xi}^k \leq 1 + \frac{C_\pi^2\sigma^2\log T}{\Delta_{i,k}^2}.$$

Then, we apply Lemma 8 again to bound the regret due to second pulls of any sub-optimal arm $i \notin \arg\max_{i\in\mathcal{K}}\mu_i^k$ in any batch k,

$$\begin{aligned}
OP(i,k) &\triangleq \sum_{t=t_k+1}^{t_{k+1}}\sum_{h=2}^{h_i^k} \mathbb{1}\left(t = t_i^k(h) \wedge \xi_t^\alpha\right)\left(\mu_\star(t) - \mu_i(t)\right)\\
&\leq C_\pi\sigma\sqrt{\left(h_{i,\xi}^k - 1\right)\log T}\\
&\leq \frac{C_\pi^2\sigma^2\log T}{\Delta_{i,k}}.
\end{aligned}$$

We apply Lemma 6 on the set of $\Upsilon_T - 1$ breakpoints and we conclude thanks to the precedent equation,

$$\begin{aligned}
\mathbb{E}\left[R_T(\pi)\right] &\leq \mathbb{E}\left[\sum_{k=0}^{\Upsilon_T-1}\sum_{i\in\mathcal{K}_k} OP(i,k)\right] + C_\pi\sigma\Upsilon_T K\sqrt{\log T} + 6KV\\
&\leq \sum_{k=0}^{\Upsilon_T-1}\sum_{i\in\mathcal{K}} \frac{C_\pi^2\sigma^2\log T}{\Delta_{i,k}} + C_\pi\sigma\Upsilon_T K\sqrt{\log T} + 6KV\,.
\end{aligned}$$

□

F Rested rotting setting

Sketch of the proof

In Lemma 9, we split the regret decomposition according to whether the overpulls has been done on the favorable event ξ_t^α or not.

In Lemma 10, we show that the part of the expected regret due to pulls under $\overline{\xi_t^\alpha}$ is bounded by a constant with respect to T for $\alpha > 4$. Indeed, while we have only trivial bounds on the quality of the pulls on these events, we can control their probabilities thanks to Proposition 2.

In Lemma 11, we show that for $h_{i,T}$ overpulls of arm i, we suffer no more than $\widetilde{\mathcal{O}}\left(\sqrt{h_{i,T}}\right)$ on the favorable event. Indeed, thanks to Lemma 4, we know that the cost of the h before last pulls is bounded by $h \cdot c(h, \delta_t) = \widetilde{\mathcal{O}}\left(\sqrt{h}\right)$.

The proof of Proposition 6 follows by noticing that $\sum_{i\in\mathcal{K}} h_{i,T} \le T$ which leads to the $\widetilde{\mathcal{O}}\left(\sqrt{KT}\right)$ rate. Indeed, thanks to the concavity of the $\sqrt{\cdot}$ and to Jensen's inequality, we find that the worst allocation is $h_{i,T} = \frac{T}{K}$.

In Lemma 12, we construct a problem-dependent bound of $h_{i,T}$ which extends the notion of gap for rotting bandits using Lemma 4.

The proof of Proposition 7 follows by plugging this bound in the result of Lemma 11.

Full proof

Let $t_i^\pi(n)$ the function such that $t_i^\pi(n) = t$ when policy π selects arm i at time t for the n-th time. We call $\mu_T^+(\pi) \triangleq \max_{i\in\mathcal{K}} \mu_i\left(N_{i,T}\right)$, *i.e.* the largest available reward for π at round T+1.

Lemma 9. *Let $h_{i,T} \triangleq |N_{i,T} - N^\star_{i,T}|$. For any policy π, the regret at round T is no bigger than*

$$R_T(\pi) \le \sum_{i\in\text{OP}} \sum_{h=0}^{h_{i,T}-1} \left[\xi^\alpha_{t_i^\pi(N^\star_{i,T}+h)}\right] \left(\mu_T^+(\pi) - \mu_i(N^\star_{i,T}+h)\right) + \sum_{t=1}^{T} \left[\overline{\xi^\alpha_t}\right] Lt.$$

We refer to the the first sum above as to A_π and to the second sum as to B.

Proof. In the rested rotting setting, we can conveniently write the regret as

$$R_T(\pi) = \sum_{i\in\mathcal{K}} \left(\sum_{n=0}^{N^\star_{i,T}-1} \mu_i(n) - \sum_{n=0}^{N^\pi_{i,T}-1} \mu_i(n) \right) \tag{62}$$

$$= \sum_{i\in\text{UP}} \sum_{n=N^\pi_{i,T}}^{N^\star_{i,T}-1} \mu_i(n) - \sum_{i\in\text{OP}} \sum_{n=N^\star_{i,T}}^{N^\pi_{i,T}-1} \mu_i(n), \tag{63}$$

where we define $\text{UP} \triangleq \left\{i \in \mathcal{K} | N^\star_{i,T} > N^\pi_{i,T}\right\}$ and likewise $\text{OP} \triangleq \left\{i \in \mathcal{K} | N^\star_{i,T} < N^\pi_{i,T}\right\}$ as the sets of arms that are respectively under-pulled and over-pulled by π with respect to the optimal policy. We consider the regret at round T.

We upper-bound all the rewards in the first double sum - the underpulls - by their maximum $\mu_T^+(\pi) \triangleq \max_{i\in\mathcal{K}} \mu_i(N^\pi_{i,T})$. Indeed, for any overpulls $\mu_i(n_i)$ (with $n_i \ge N^\pi_{i,T}$), we have that

$$\mu_i(n_i) \le \mu_i(N^\pi_{i,T}) \le \mu_T^+(\pi) \triangleq \max_{i\in\mathcal{K}} \mu_i(N^\pi_{i,T}),$$

where the first inequality follows by the non-increasing property of μ_is; and the second by the definition of the maximum operator. Second, we notice that there are as many underpulls than overpulls (terms of the second double sum) because there both policies π and $\pi^\star$ pull T arms. Notice that this does *not* mean that for each arm i, the number of overpulls equals to the number of underpulls, which cannot happen anyway since an arm cannot be simultaneously underpulled and overpulled. Therefore, we keep only the second double sum,

$$R_T(\pi) \le \sum_{i\in\text{OP}} \sum_{n=N^\star_{i,T}}^{N^\pi_{i,T}-1} \left(\mu_T^+(\pi) - \mu_i(n)\right). \tag{64}$$

Then, we need to separate overpulls that are done under ξ^α_t and under $\overline{\xi^\alpha_t}$. We introduce $t_i^\pi(n)$, the round at which π pulls arm i for the n-th time. We now make the round at which each overpull occurs explicit,

$$\begin{aligned} R_T(\pi) &\leq \sum_{i\in\textsc{op}} \sum_{h=0}^{h_{i,\,T}-1} \sum_{t=1}^{T} \left[t_i^\pi\left(N^\star_{i,\,T}+h\right)=t\right]\left(\mu_T^+(\pi)-\mu_i(N^\star_{i,\,T}+h)\right) \\ &\leq \underbrace{\sum_{i\in\textsc{op}} \sum_{h=0}^{h_{i,\,T}-1} \sum_{t=1}^{T} \left[t_i^\pi\left(N^\star_{i,\,T}+h\right)=t\wedge\xi_t^\alpha\right]\left(\mu_T^+(\pi)-\mu_i(N^\star_{i,\,T}+h)\right)}_{A_\pi} \\ &+ \underbrace{\sum_{i\in\textsc{op}} \sum_{h=0}^{h_{i,\,T}-1} \sum_{t=1}^{T} \left[t_i^\pi\left(N^\star_{i,\,T}+h\right)=t\wedge\overline{\xi_t^\alpha}\right]\left(\mu_T^+(\pi)-\mu_i(N^\star_{i,\,T}+h)\right)}_{B}. \end{aligned}$$

For the analysis of the pulls done under ξ_t^α we do not need to know at which round it was done. Therefore,

$$A_\pi \leq \sum_{i\in\textsc{op}} \sum_{h=0}^{h_{i,\,T}-1} \left[\xi^\alpha_{t(N^\star_{i,\,t}+h)}\right]\left(\mu_T^+(\pi)-\mu_i(N^\star_{i,\,T}+h)\right).$$

For `FEWA` or `RAW-UCB`, it is not easy to directly guarantee the low probability of overpulls (the second sum). Thus, we upper-bound the regret of each overpull at round t under $\overline{\xi_t^\alpha}$ by its maximum value Lt. While this is done to ease `FEWA` analysis, this is valid for any policy π. Then, noticing that we can have at most 1 overpull per round t, i.e., $\sum_{i\in\textsc{op}}\sum_{h=0}^{h_{i,\,T}-1}\left[t_i^\pi\left(N^\star_{i,\,T}+h\right)=t\right]\leq 1$, we get

$$B \leq \sum_{t=1}^{T}\left[\overline{\xi_t^\alpha}\right]Lt\left(\sum_{i\in\textsc{op}}\sum_{h=0}^{h_{i,\,T}-1}\left[t_i^\pi\left(N^\star_{i,\,T}+h\right)=t\right]\right) \leq \sum_{t=1}^{T}\left[\overline{\xi_t^\alpha}\right]Lt.$$

Therefore, we conclude that

$$R_T(\pi) \leq \underbrace{\sum_{i\in\textsc{op}} \sum_{h=0}^{h_{i,\,T}-1} \left[\xi^\alpha_{t_i^\pi(N^\star_{i,\,t}+h)}\right]\left(\mu_T^+(\pi)-\mu_i(N^\star_{i,\,T}+h)\right)}_{A_\pi} + \underbrace{\sum_{t=1}^{T}\left[\overline{\xi_t^\alpha}\right]Lt}_{B}.$$

□

Lemma 10. *Let $\zeta(x)=\sum_n n^{-x}$. Thus, with $\delta_t=2t^{-\alpha}$ and $\alpha>4$, we can use Proposition 2 and get*

$$\mathbb{E}\left[B\right] \triangleq \sum_{t=1}^{T} p\left(\overline{\xi_t^\alpha}\right)Lt \leq \sum_{t=1}^{T} KLt^{3-\alpha} \leq KL\zeta(\alpha-3).$$

In particular, for $\alpha\geq 5$, we have :

$$\mathbb{E}\left[B\right] \leq KL\zeta(2) \leq 2KL < 5KL.$$

Thanks to Proposition 11, we can prove a similar bound on $B_2 \triangleq \sum_{t=1}^T\left[\overline{\xi_{t,\,2}^\alpha}\right]Lt$,

$$\mathbb{E}\left[B_2\right] \leq 6KL\zeta(\alpha-2),\ \text{ i.e., } \textit{for } \alpha\geq 4,\ \mathbb{E}\left[B_2\right]\leq 5KL.$$

Lemma 11. *We define $h^\xi_{i,\,T} \triangleq \max\left\{h\leq h_{i,\,T} \middle|\ \xi^\alpha_{t_i^\pi(N^\star_{i,\,t}+h)}\right\}$, the largest number of overpulls of arm i pulled under ξ_t^α at round $t=t_i^\pi(N^\star_{i,\,t}+h^\xi_{i,\,T})\leq T$. We also define $\textsc{op}_\xi \triangleq \left\{i\in\textsc{op}\middle|\ h^\xi_{i,\,T}\geq 1\right\}$. For policy $\pi\in\{\pi_\mathrm{R},\pi_\mathrm{F}\}$ with parameter α, A_π defined in Lemma 9 is upper-bounded by*

$$\begin{aligned} A_\pi &\triangleq \sum_{i\in\textsc{op}} \sum_{h=0}^{h_{i,\,T}-1} \left[\xi^\alpha_{t_i^\pi(N^\star_{i,\,T}+h)}\right]\left(\mu_T^+(\pi)-\mu_i(N^\star_{i,\,T}+h)\right) \\ &\leq \sum_{i\in\textsc{op}_\xi}\left(C_\pi\sigma\sqrt{\left(h^\xi_{i,\,T}-1\right)\log\left(T\right)}+C_\pi\sigma\sqrt{\log\left(T\right)}+L\right). \end{aligned}$$

Proof. First, we define $h^{\xi}_{i,\,T} \triangleq \max\left\{h \leq h_{i,\,T} \middle| \; \xi^{\alpha}_{t^{\pi}_{i}(N^{\star}_{i,\,t}+h)}\right\}$, the largest number of overpulls of arm i pulled at round $t_i \triangleq t^{\pi}_{i}(N^{\star}_{i,\,t} + h^{\xi}_{i,\,T}) \leq T$ under ξ^{α}_{t}. Now, we upper-bound A_π by including all the overpulls of arm i until the $h^{\xi}_{i,\,T}$-th overpull, even the ones under $\overline{\xi^{\alpha}_{t}}$,

$$
\begin{aligned}
A_\pi &\triangleq \sum_{i \in \textsc{op}} \sum_{h=0}^{h_{i,\,T}-1} \left[\xi^{\alpha}_{t^{\pi}_{i}(N^{\star}_{i,\,t}+h)}\right] \left(\mu^{+}_{T}(\pi) - \mu_i(N^{\star}_{i,\,T}+h)\right) \\
&\leq \sum_{i \in \textsc{op}_\xi} \sum_{h=0}^{h^{\xi}_{i,\,T}} \left(\mu^{+}_{T}(\pi) - \mu_i(N^{\star}_{i,\,T}+h)\right),
\end{aligned}
$$

where $\textsc{op}_\xi \triangleq \left\{i \in \textsc{op} \middle| \; h^{\xi}_{i,\,T} \geq 1\right\}$. We can therefore split the second sum of $h^{\xi}_{i,\,T}$ term above into two parts. The first part corresponds to the first $h^{\xi}_{i,\,T} - 1$ (possibly zero) terms (overpulling differences) and the second part to the last $(h^{\xi}_{i,\,T} - 1)$-th one. Recalling that at round t_i, arm i was selected under $\xi^{\alpha}_{t_i}$, we apply Lemma 4 to bound the regret caused by previous overpulls of i (possibly none),

$$
\begin{aligned}
A_\pi &\leq \sum_{i \in \textsc{op}_\xi} \mu^{+}_{T}(\pi) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,\,T} - 1\right) + \frac{C_\pi}{\sqrt{2\alpha}}\left(h^{\xi}_{i,\,T} - 1\right) c\left(h^{\xi}_{i,\,T} - 1, \delta_{t_i}\right) && (65) \\
&\leq \sum_{i \in \textsc{op}_\xi} \mu^{+}_{T}(\pi) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,\,T} - 1\right) + \frac{C_\pi}{\sqrt{2\alpha}}\left(h^{\xi}_{i,\,T} - 1\right) c\left(h^{\xi}_{i,\,T} - 1, \delta_{T}\right) && (66) \\
&\leq \sum_{i \in \textsc{op}_\xi} \mu^{+}_{T}(\pi) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,\,T} - 1\right) + C_\pi \sigma \sqrt{\left(h^{\xi}_{i,\,T} - 1\right) \log(T)}, && (67)
\end{aligned}
$$

The second inequality is obtained because δ_t is decreasing and $c(.,\delta)$ is decreasing as well. The last inequality is the definition of confidence interval in Proposition 2. If $N^{\star}_{i,\,T} = 0$ and $h^{\xi}_{i,\,T} = 1$ then

$$
\mu^{+}_{T}(\pi) - \mu_i(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 1) = \mu^{+}_{T}(\pi) - \mu_i(0) \leq L,
$$

since $\mu^{+}_{T}(\pi) \leq \max_{j \in \mathcal{K}} \mu_j(0)$ and $\max_{j \in \mathcal{K}} \mu_j(0) - \mu_i(0) \leq L$ by the definition of L (Eq. 10). Otherwise, we can decompose

$$
\begin{aligned}
\mu^{+}_{T}(\pi) - \mu_i(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 1) = &\underbrace{\mu^{+}_{T}(\pi) - \mu_i(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 2)}_{A_1} \\
&+ \underbrace{\mu_i(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 2) - \mu_i(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 1)}_{A_2}.
\end{aligned}
$$

For term A_1, since this $h^{\xi}_{i,\,T}$-th overpull is done under $\xi^{\alpha}_{t_i}$, by Lemma 4 we have that

$$
A_1 = \mu^{+}_{T}(\pi) - \overline{\mu}^{1}_{i}(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 1) \leq 1c(1,\delta_{t_i}) \leq 2c(1,\delta_T) \leq C_\pi \sigma \sqrt{\log(T)}.
$$

The second difference, $A_2 = \mu_i(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 2) - \mu_i(N^{\star}_{i,\,T} + h^{\xi}_{i,\,T} - 1)$ cannot exceed L by definition (Eq. 10), the maximum decay in one round is bounded. Therefore, we further upper-bound Equation 67 as

$$
A_\pi \leq \sum_{i \in \textsc{op}_\xi} \left(C_\pi \sigma \sqrt{\left(h^{\xi}_{i,\,T} - 1\right) \log(T)} + C_\pi \sigma \sqrt{\log(T)} + L\right). \tag{68}
$$

□

Proposition 6 (gap-free bound)**.** *Let $\pi \in \{\pi_{\mathrm{F}}, \pi_{\mathrm{R}}\}$ tuned with $\alpha \geq 5$ or $\pi \in \{\pi_{\mathrm{EF}}, \pi_{\mathrm{ER}}\}$ tuned with $\alpha \geq 4$ and $m = 2$. For any rotting bandit scenario with means $\{\mu_i\}_i$ satisfying Assumption 1 with bounded decay L and any time horizon T, π suffers an expected regret,*

$$
\mathbb{E}\left[R_T(\pi)\right] \leq C_\pi \sigma \sqrt{\log(T)} \left(\sqrt{KT} + K\right) + 6KL.
$$

Proof. In Lemma 9, we split the regret in two parts. The first one B corresponds to the regret due to unfavorable events $\overline{\xi_t^\alpha}$. We do not derive any guarantee of our algorithms on these events but their probabilities can be controlled thanks to parameter α. Hence, for $\alpha > 4$, we show in Lemma 10 that the part of the expected regret due to unfavorable events can be bounded by a constant w.r.t. T. Yet, we choose $\alpha \geq 5$ to have a small constant.

The second one A_π corresponds to the regret due to favorable events ξ_t^α which can be bounded for our two algorithms (`FEWA` and `RAW-UCB`) thanks to Lemma 11. In order to get a problem-independent upper bound, we need to replace $h_{i,T}^\xi$ by a problem-independent quantity. Starting from Lemma 11,

$$A_\pi \leq \sum_{i \in \textsc{op}_\xi} \left(C_\pi \sigma \sqrt{\left(h_{i,T}^\xi - 1\right) \log(T)} + C_\pi \sigma \sqrt{\log(T)} + L \right).$$

Since $\textsc{op}_\xi \subseteq \textsc{op}$, we can upper-bound the number of terms in the above sum by K. Next, the total number of overpulls $\sum_{i \in \textsc{op}} h_{i,T}$ cannot exceed T. As square-root function is concave we can use Jensen's inequality. Moreover, we can deduce that the worst allocation of overpulls is the uniform one, i.e., $h_{i,T} = T/K$,

$$\begin{aligned} A_\pi &\leq K(C_\pi \sigma \sqrt{\log(T)} + L) + C_\pi \sigma \sqrt{\log(T)} \sum_{i \in \textsc{op}} \sqrt{(h_{i,T} - 1)} && (69) \\ &\leq K(C_\pi \sigma \sqrt{\log(T)} + L) + C_\pi \sigma \sqrt{KT \log(T)}. && (70) \end{aligned}$$

Therefore, using Lemma 9 together with Equations 70 and Lemma 10, we bound the total expected regret as

$$\mathbb{E}[R_T(\pi)] \leq C_\pi \sigma \sqrt{\log(T)} \left(\sqrt{KT} + K\right) + 6KL\cdot \qquad (71)$$

□

Lemma 12. *We define the smallest reward gathered by the optimal policy μ_T^- and the gap of the h first overpulls of arm i with respect to that value $\Delta_{i,h}$.*

$$\begin{aligned} &\mu_T^- \triangleq \min_{i \in \mathcal{K}^\star} \mu_i \left(N_{i,T}^\star - 1\right) \;\; \textit{with } \mathcal{K}^\star \triangleq \left\{ i \in \mathcal{K} | N_{i,T}^\star \geq 1 \right\}, \\ &\Delta_{i,h} \triangleq \mu_T^- - \overline{\mu}_i^h \left(N_{i,t}^\star + h\right). \end{aligned}$$

$h_{i,T}^\xi$ *defined in Lemma 9 is upper-bounded by a problem-dependent quantity,*

$$h_{i,T}^\xi \leq h_{i,T}^+ \triangleq \max \left\{ h \leq T |\; h \leq 1 + \frac{C_\pi^2 \sigma^2 \log(T)}{\Delta_{i,h-1}^2} \right\} \leq 1 + \frac{C_\pi^2 \sigma^2 \log(T)}{\Delta_{i,h_{i,T}^+ - 1}^2}.$$

Proof. We want to bound $h_{i,T}^\xi$ with a problem dependent quantity $h_{i,T}^+$. We remind the reader that for arm i at round T, the $h_{i,T}^\xi$-th overpull is pulled under $\xi_{t_i}^\alpha$ at round t_i. Therefore, Lemma 4 applies and we have

$$\begin{aligned} \overline{\mu}_i^{h_{i,T}^\xi - 1} \left(N_{i,T}^\star + h_{i,T}^\xi - 1\right) &\geq \mu_T^+(\pi) - \frac{C_\pi}{\sqrt{2\alpha}} c\left(h_{i,T}^\xi - 1, \delta_{t_i}\right) \\ &\geq \mu_T^+(\pi) - \frac{C_\pi}{\sqrt{2\alpha}} c\left(h_{i,T}^\xi - 1, \delta_T\right) \\ &\geq \mu_T^+(\pi) - C_\pi \sigma \sqrt{\frac{\log(T)}{h_{i,T}^\xi - 1}}, \end{aligned}$$

Hence, we have that

$$h_{i,T}^\xi \leq 1 + \frac{C_\pi^2 \sigma^2 \log(T)}{\left(\mu_T^+(\pi) - \overline{\mu}_i^{h_{i,T}^\xi - 1}\left(N_{i,T}^\star + h_{i,T}^\xi - 1\right)\right)^2}. \qquad (72)$$

Yet, this upper-bound still depends on random quantities such as $\mu_T^+(\pi)$ or $h_{i,T}^{\xi}$ on the denominator. Consider the smallest value collected by the optimal policy,

$$\mu_T^- \triangleq \min_{i\in\mathcal{K}^\star} \mu_i\left(N_{i,T}^\star - 1\right) \text{ with } \mathcal{K}^\star \triangleq \left\{i \in \mathcal{K} | N_{i,T}^\star \geq 1\right\}.$$

It is the T-th largest value among the KT possible ones. Since $\overline{\mu}_i^{h_{i,T}^{\xi}-1}\left(N_{i,T}^\star + h_{i,T}^{\xi} - 1\right)$ is an average of overpulls value, which are all smaller or equal to μ_T^-, we have

$$\mu_T^- \geq \overline{\mu}_i^{h_{i,T}^{\xi}-1}\left(N_{i,T}^\star + h_{i,T}^{\xi} - 1\right).$$

Moreover, $\mu_T^- > \mu_T^+(\pi)$ implies that the regret is 0. Indeed, in that case $\mu_T^+(\pi)$ - the pull with the largest value among the remaining values at the end of the game for π - is *strictly smaller* than μ_T^- - the T-th largest reward sample. Therefore, π has collected the T largest value and has zero regret. Hence, we focus on the case $\mu_T^- \leq \mu_T^+(\pi)$, for which the regret may not be zero. In that case, we can upper-bound the RHS term Equation 72 by replacing the random quantity $\mu_T^+(\pi)$ by the smaller quantity μ_T^-. Hence,

$$h_{i,T}^{\xi} \leq 1 + \frac{C_\pi^2\sigma^2\log(T)}{\left(\mu_T^+(\pi) - \overline{\mu}_i^{h_{i,T}^{\xi}-1}\left(N_{i,T}^\star + h_{i,T}^{\xi} - 1\right)\right)^2} \leq 1 + \frac{C_\pi^2\sigma^2\log(T)}{\Delta^2_{i,h_{i,T}^{\xi}-1}},$$

with $\Delta_{i,h} \triangleq \mu_T^- - \overline{\mu}_i^h\left(N_{i,t}^\star + h\right)$, the difference between the lowest mean value of the arm pulled by $\pi^\star$ and the average of the h first overpulls of arm i. Yet, this self-bounding property of $h_{i,T}^{\xi}$ is not a proper problem-dependent upper bound. We will consider the largest h which satisfies this self-bounding property,

$$h_{i,T}^+ \triangleq \max\left\{h \leq T |\ h \leq 1 + \frac{C_\pi^2\sigma^2\log(T)}{\Delta^2_{i,h-1}}\right\}.$$

We have that,

$$h_{i,T}^{\xi} \leq h_{i,T}^+ \leq 1 + \frac{C_\pi^2\sigma^2\log(T)}{\Delta^2_{i,h_{i,T}^+-1}}.$$

□

Proposition 7 (gap-dependent bound). *$\pi \in \{\pi_\mathrm{F}, \pi_\mathrm{R}\}$ tuned with $\alpha \geq 5$ (or $\pi \in \{\pi_\mathrm{EF}, \pi_\mathrm{ER}\}$ tuned with $\alpha \geq 4$ and $m = 2$) suffers an expected regret,*

$$\mathbb{E}\left[R_T(\pi)\right] \leq \sum_{i\in\mathcal{K}}\left(\frac{C_\pi^2\sigma^2\log(T)}{\Delta_{i,h_{i,T}^+-1}} + C_\pi\sigma\sqrt{\log(T)} + 6L\right)$$

with $h_{i,T}^+ \triangleq \max\left\{h \leq 1 + \frac{C_\pi^2\sigma^2\log T}{\Delta^2_{i,h-1}}\right\}$, *and the pseudo-gap*

$$\Delta_{i,h} \triangleq \min_{j\in\mathcal{K}} \mu_j\left(N_{j,T}^\star - 1\right) - \overline{\mu}_i^h\left(N_{i,T}^\star + h\right).$$

Proof. We use Lemmas 11 and Lemma 12 to bound A_π (see Lemma 9). Indeed, since the square-root function is increasing, we can upper-bound the result in Lemma 11 by replacing $h_{i,T}^{\xi}$ by its upper bound in Lemma 12

$$\begin{aligned} A_\pi &\leq \sum_{i\in\mathrm{OP}_\xi}\left(C_\pi\sigma\sqrt{\log(T)}\left(1 + \sqrt{h_{i,T}^+ - 1}\right) + L\right) \\ &\leq \sum_{i\in\mathrm{OP}_\xi}\left(C_\pi\sigma\sqrt{\log(T)}\left(1 + \frac{C_\pi\sigma\sqrt{\log(T)}}{\Delta_{i,h_{i,T}^+-1}}\right) + L\right). \end{aligned}$$

Notice that the quantity $\mathrm{OP}_\xi \subset \mathcal{K}$. Therefore, we have

$$A_\pi \leq \sum_{i\in\mathcal{K}}\left(\frac{C_\pi^2\sigma^2\log(T)}{\Delta_{i,h_{i,T}^+-1}} + C_\pi\sigma\sqrt{\log(T)} + L\right). \tag{73}$$

Using Lemmas 9, 10, and Equation 73 we get

$$\begin{aligned}
\mathbb{E}\left[R_T(\pi)\right] &= \mathbb{E}\left[A_\pi\right] + \mathbb{E}\left[B\right] \\
&\leq \sum_{i\in\mathcal{K}} \left(\frac{C_\pi^2 \sigma^2 \log(T)}{\Delta_{i,h_{i,T}^+ - 1}} + C_\pi \sigma \sqrt{\log(T)} + L \right) + 5KL \\
&\leq \sum_{i\in\mathcal{K}} \left(\frac{C_\pi^2 \sigma^2 \log(T)}{\Delta_{i,h_{i,T}^+ - 1}} + C_\pi \sigma \sqrt{\log(T)} + 6L \right) \cdot
\end{aligned}$$

□

G Full experiments

The code of all our experiments can be found on SMPyBandits (Besson, 2018), an open-source bandits package in Python. The goal of these experiments is to perform an exhaustive benchmark of non-stationary algorithms which might be able to perform well in both rested and restless rotting setups in an agnostic way (*i.e.* with the same tuning).

Algorithms and parameters. We include UCB and FEWA (Seznec et al., 2019), the only algorithms which got known regret bounds in both setups. We include two versions of each algorithm: with the theoretical tuning $\alpha = 4$; and with the empirical tuning $\alpha_\mathrm{R} = 1.5$ and $\alpha_\mathrm{F} = 0.06$. These two values are selected by grid-search on the rested benchmark. This benchmark has 30 different problems (for different L) but this is not a problem as the best tuning of α is the same for all the considered problem. In the restless setting, we replace RAW-UCB and FEWA by their efficient versions because of the longer horizon.

We also include Exp3.S (Auer et al., 2003), an algorithm which was designed for the very general adversarial bandits problem against switching experts. As explained in the introduction, tuned Exp3.S reaches the minimax optimal rate in all the presented restless setup. Yet, it is unclear if it is able to learn in the rested rotting bandits problem. We use the theoretical tuning which requires the knowledge of T and V_T.

We also include GLR-UCB (Besson and Kaufmann, 2019). This algorithm has two parameters : a confidence level δ for its change-point detector and an active exploration rate α. We set α to zero. Indeed, the active exploration of change-detection algorithms is only useful in the increasing case (as argued by Cao et al. (2019)). We tune δ by its theoretical value, which requires the knowledge of T. Last, we only restart the history of the changed arm as our setup do not assume that all the rewards change simultaneously. For fair comparison, we only use the subgaussian version of the algorithm. Indeed, KL-UCB indexes are expensive to compute. Instead, for all the confidence bound algorithms, we rather tune $\sigma^2 = 1$ in the rested benchmark and $\sigma^2 = 0.29$ in the restless benchmark (the variance of a binomial $\mathcal{B}(10, 0.03)$).

We do not include SWA (Levine et al., 2017) which was shown to be less consistent than FEWA (Seznec et al., 2019) on rested rotting bandits. We do not include SW-UCB and D-UCB as they were shown to be unable to learn in the rested setting (Levine et al., 2017; Seznec et al., 2019). We also do not include CUSUM-UCB (Liu et al., 2018) and M-UCB (Cao et al., 2019), as 1) they were shown to under-perform against GLR-UCB (Besson and Kaufmann, 2019); and 2) their change-detector is harder to tune.

G.1 Simulated benchmark for rested bandits

Setup. We use the two-arm benchmark of Seznec et al. (2019). Arms are gaussians with fix variance $\sigma = 1$ and rested rotting mean. The first arm has a constant mean 0 while the second arm abruptly switches from $+\frac{L}{2}$ to $-\frac{L}{2}$ at $t = \frac{T}{4} = 2500$. Several values of L are investigated between 10^{-3} and 10.

Result : RAW-UCB vs FEWA. We compare RAW-UCB and FEWA both for theoretical value $\alpha = 4$ and tuned values $\alpha_\mathrm{R} = 1.5$ and $\alpha_F = 0.06$ (selected by grid-search). For theoretical tuning, we see in Figure 2 (left), that RAW-UCB outperforms FEWA on all sizes of decays by a factor ~ 4 which is predicted by our theory. Indeed, there is also a factor 4 between the two problem-dependent upper-bounds. Surprisingly, for empirical tuning, the average performances of the two algorithms are much closer. We also notice that there is a larger variance in FEWA's result

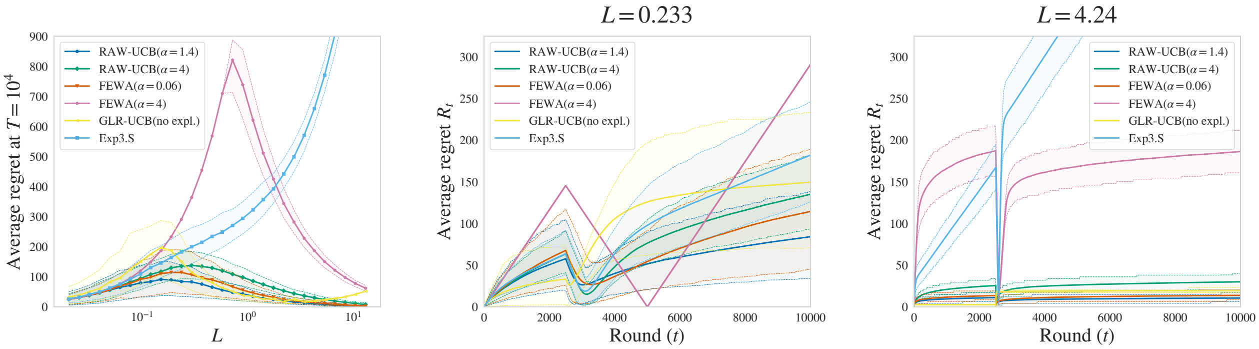


Figure 2: **Left:** Regret at the end of the game for different values of L. **Middle, Right:** Regret across time for two values of L. Average over 2000 runs. We highlight the $[10\%, 90\%]$ confidence region.

compared to RAW-UCB. This is not surprising because we had to drastically reduce the confidence bounds to make FEWA practical. It means that empirical FEWA filters arms based only on a handful of samples. This bet leads to either very good runs or very bad runs. Last, Figure 2 (middle, right) shows that RAW-UCB outperforms FEWA at any time T, both on easy and difficult problems.

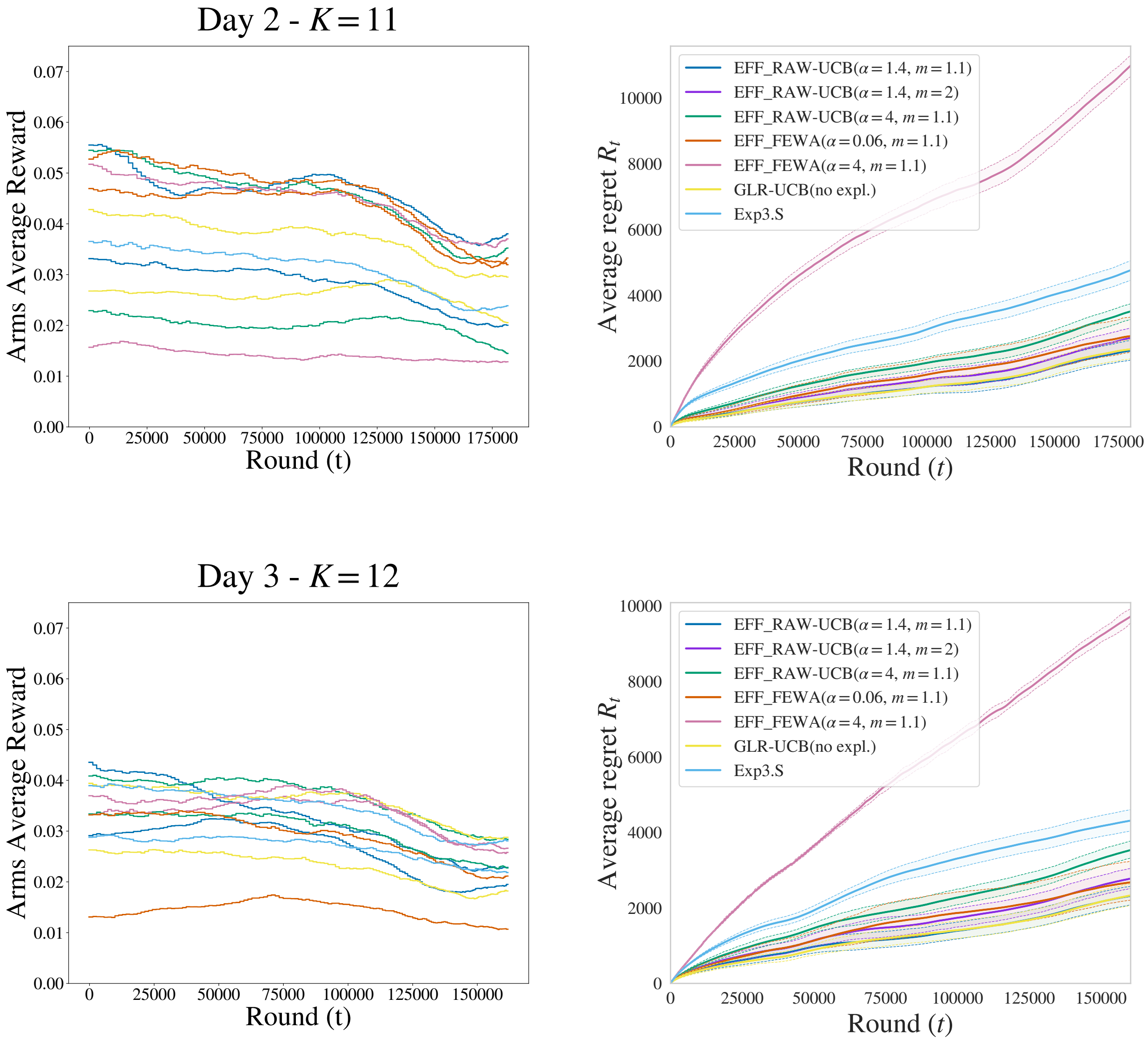
Overall, our experiment suggests that RAW-UCB has better expected and high-probability performance than FEWA on rested problems. Moreover, our analysis reduces the gap between theory and practice. Indeed, RAW-UCB practical confidence bounds are reduced compared to theoretical value by a factor $\sqrt{4/1.4} = 1.7$ while FEWA's are reduced by a factor $\sqrt{4/0.06} = 8.2$. Note that the empirical tuning $\delta_T = T^{-1.4}$ is very close to asymptotic optimal tuning of UCB: $\delta_T = T^{-1}\log(T)^{-2} \sim T^{-1.48}$ for $T = 10^4$. It suggests that RAW-UCB might not need to use larger confidence bands than UCB for stationary bandits.

Result : Restless algorithms. Exp3.S shows reasonable performance for small L and very bad performance for large L. Indeed, Exp3.S suffers from the fact that it pulls any arm with a probability at least $\sqrt{T}^{-1}$. When the cost of a single mistake is big (large L), it increases the regret. When the distance between arm is small (small L), all the consistent policies do $\sim T$ number of mistakes. Hence, the $\sqrt{T}^{-1}$ exploration rate is not a problem here. Combined with the observation of Levine et al. (2017) and Seznec et al. (2019), we can conclude that passive forgetting and active random exploration leads to linear regret rate in rested rotting bandits.

This is why we cancel the random exploration for GLR-UCB. GLR-UCB use an active forgetting mechanism based on change-point detection. While it has been designed for restless bandits, it gives surprisingly good result on this rested benchmark. For $L < 10^{-1}$, GLR-UCB shows worse regret than FEWA $(\alpha = 4)$ which is equivalent to round-robin for small L. This is because the change is too small to be detected. Hence, the algorithm uses biased sample to compute the suboptimal arm's UCB after the break point. In this region, there is also large regret deviation. For $L \in [0.1, 4]$, GLR-UCB performs very well. Indeed, it can detect the change-point and then run the optimal KL-UCB subroutine for the remaining rounds (on which reward is stationary). For $L > 4$, GLR-UCB have worse performance than RAW-UCB. Indeed, when an arm gets abruptly worse, 1) we detect the change-point; 2) we restart the arm's history which triggers additional exploratory pulls. This two steps mechanism require more pulls at the break-point than RAW-UCB. However, we can see on Figure 2 (right) that after the first few pulls GLR-UCB pulls the sub-optimal arm at an optimal $\log T$ rate. This benchmark reveals that GLR-UCB with local restarts and no random exploration may be able to learn in the rested rotting setting, in particular when the decay is limited compared to the noise.

G.2 Real world Yahoo! experiment

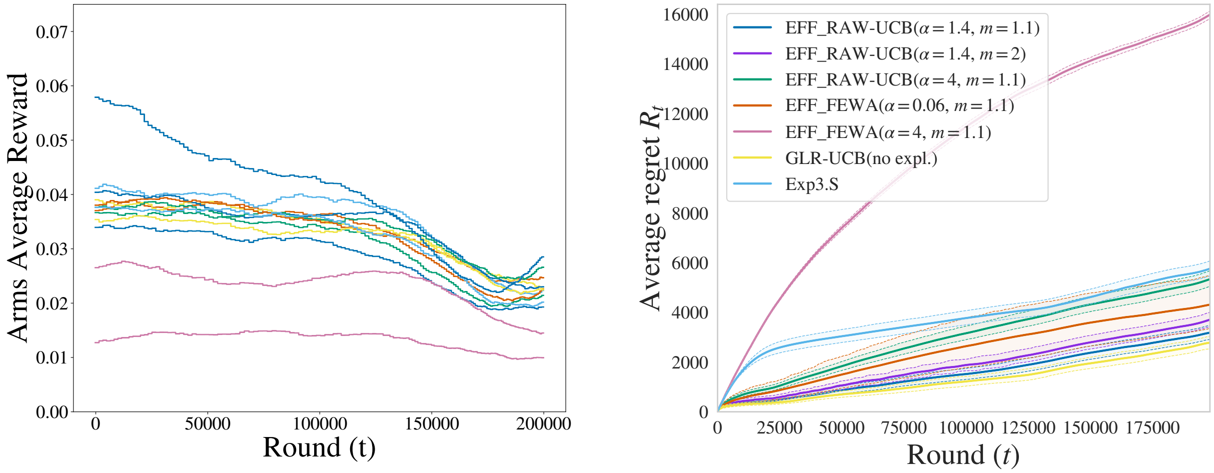
Figure 3: **Left:** reward functions on from the Yahoo! dataset
Right: average regret of policies over 500 runs



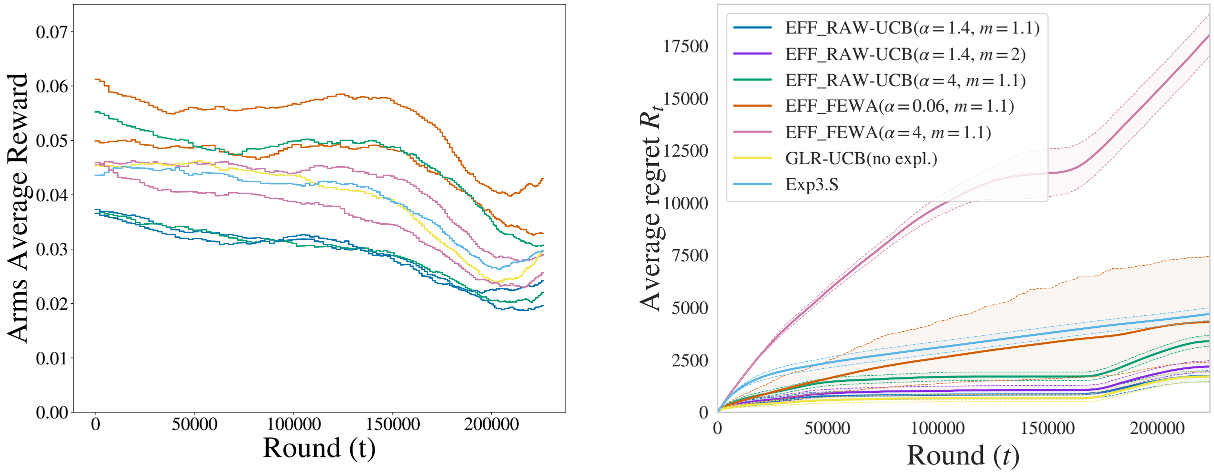
Day	EFF-RAW-UCB	EFF-RAW-UCB	EFF-FEWA	EFF-FEWA	EFF-FEWA	Exp3.S	GLR-UCB
(T)	($\alpha = 1.4, m = 1.1$)	($\alpha = 1.4, m = 2$)	($\alpha = 4, m = 1.1$)	($\alpha = 0.06$)	($\alpha = 4$)		
2	67	35	65	143	337	56	560
3	66	33	65	175	308	53	613
4	90	43	90	223	391	67	683
5	86	47	88	159	473	77	2421
6	91	46	91	183	487	75	707
7	74	41	74	115	380	69	1529
8	88	44	89	193	428	71	957
9	64	34	63	116	341	55	971

Table 1: Average computational time in seconds for each algorithm in each experiment.

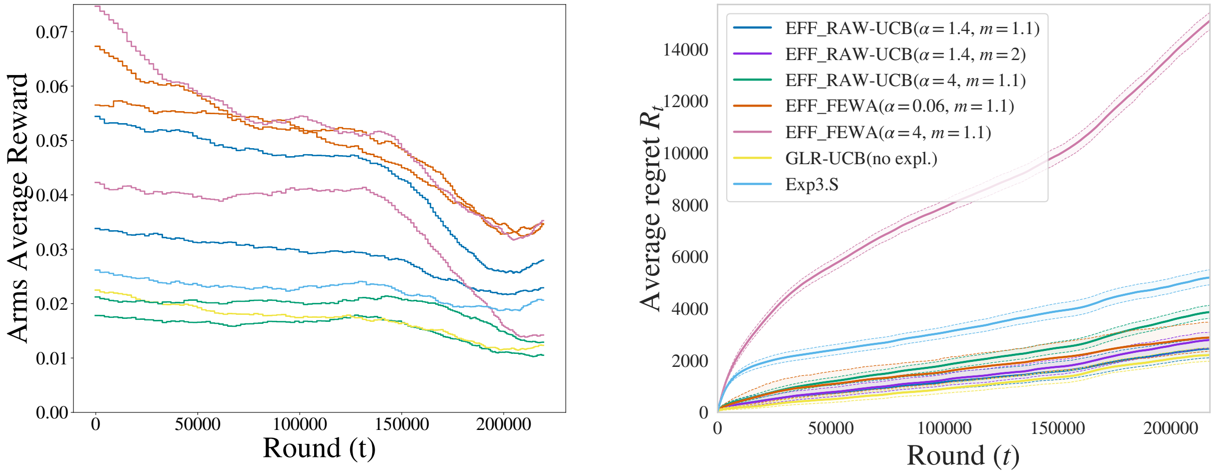
Day 4 - $K = 13$

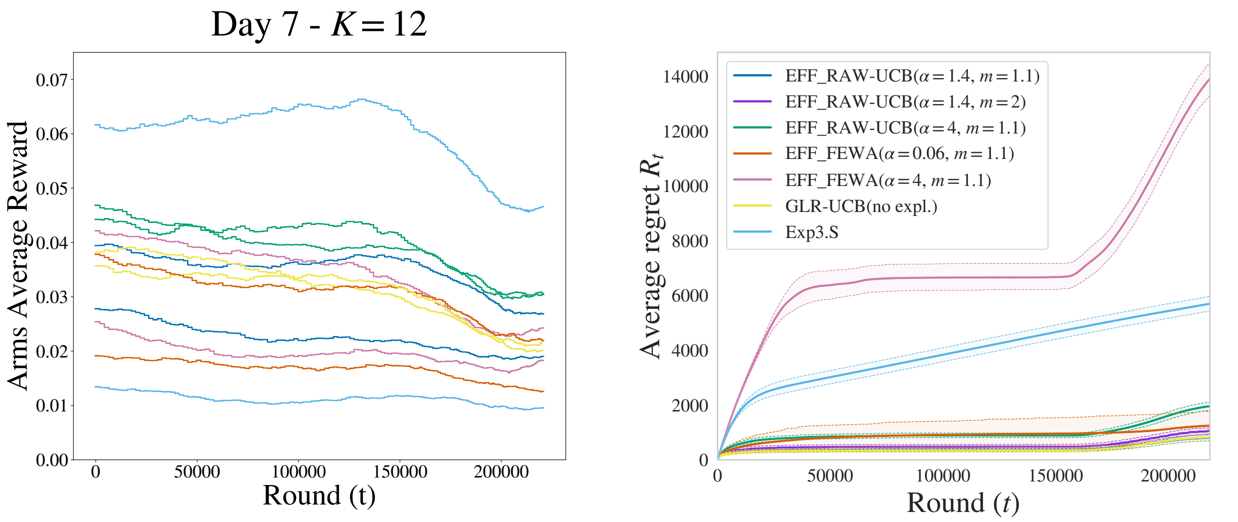


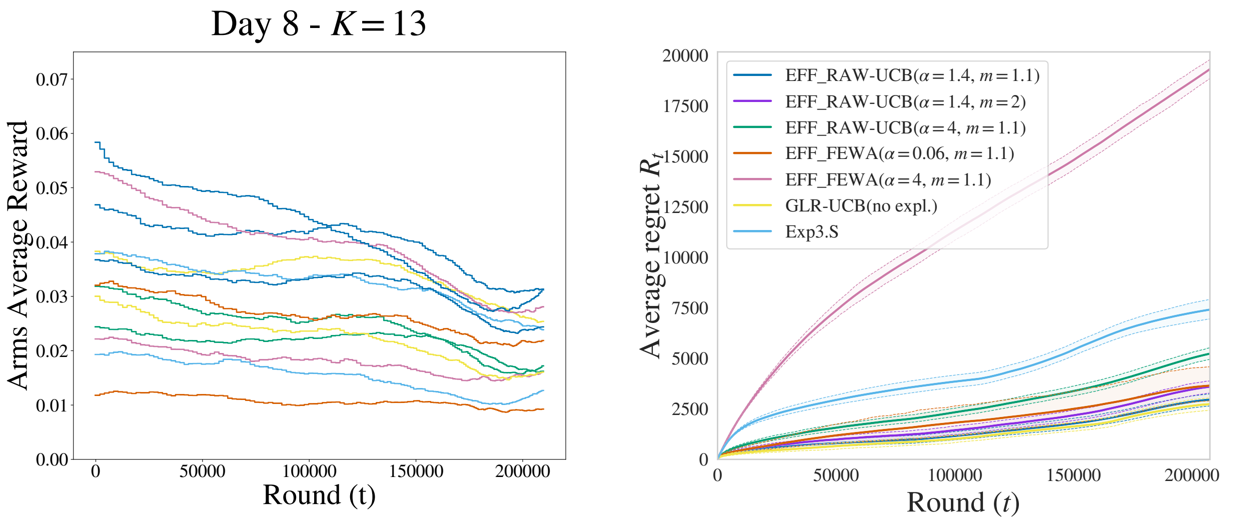
Day 5 - $K = 10$

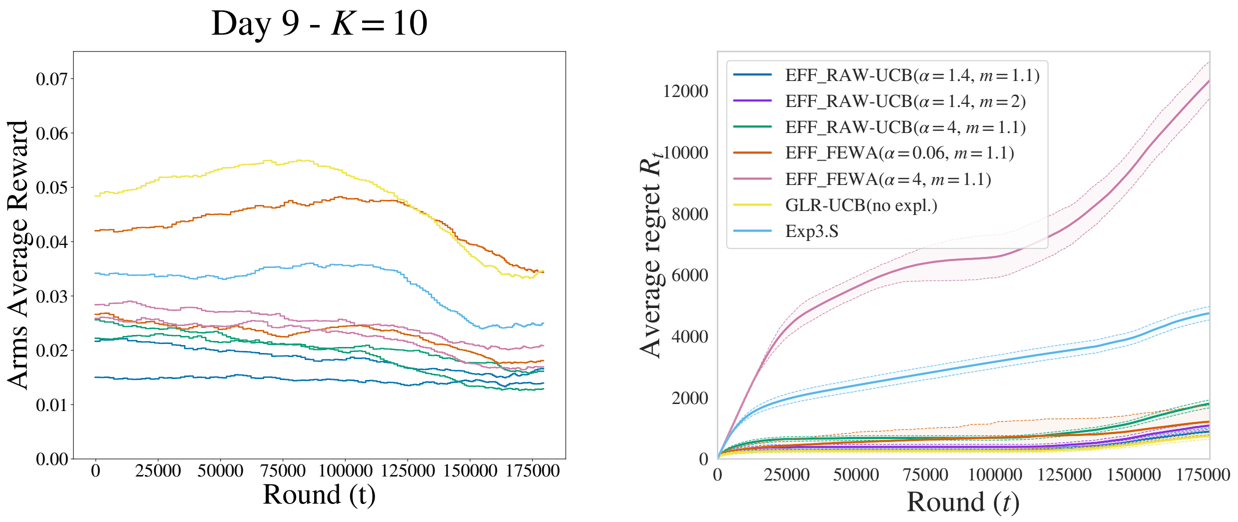


Day 6 - $K = 10$




Day 7 - $K = 12$
Arms Average Reward
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
50000
100000
150000
200000
Round (t)
Average regret R_t
0
2000
4000
6000
8000
10000
12000
14000
EFF_RAW-UCB($\alpha = 1.4$, $m = 1.1$)
EFF_RAW-UCB($\alpha = 1.4$, $m = 2$)
EFF_RAW-UCB($\alpha = 4$, $m = 1.1$)
EFF_FEWA($\alpha = 0.06$, $m = 1.1$)
EFF_FEWA($\alpha = 4$, $m = 1.1$)
GLR-UCB(no expl.)
Exp3.S
0
50000
100000
150000
200000
Round (t)


Day 8 - $K = 13$
Arms Average Reward
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
50000
100000
150000
200000
Round (t)
Average regret R_t
0
2500
5000
7500
10000
12500
15000
17500
20000
EFF_RAW-UCB($\alpha = 1.4$, $m = 1.1$)
EFF_RAW-UCB($\alpha = 1.4$, $m = 2$)
EFF_RAW-UCB($\alpha = 4$, $m = 1.1$)
EFF_FEWA($\alpha = 0.06$, $m = 1.1$)
EFF_FEWA($\alpha = 4$, $m = 1.1$)
GLR-UCB(no expl.)
Exp3.S
0
50000
100000
150000
200000
Round (t)


Day 9 - $K = 10$
Arms Average Reward
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
25000
50000
75000
100000
125000
150000
175000
Round (t)
Average regret R_t
0
2000
4000
6000
8000
10000
12000
EFF_RAW-UCB($\alpha = 1.4$, $m = 1.1$)
EFF_RAW-UCB($\alpha = 1.4$, $m = 2$)
EFF_RAW-UCB($\alpha = 4$, $m = 1.1$)
EFF_FEWA($\alpha = 0.06$, $m = 1.1$)
EFF_FEWA($\alpha = 4$, $m = 1.1$)
GLR-UCB(no expl.)
Exp3.S
0
25000
50000
75000
100000
125000
150000
175000
Round (t)