Rotting bandits are not harder than stochastic ones

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, Michal Valko

Rotting bandits are no harder than stochastic ones

Julien Seznec[1,2], Andrea Locatelli[3], Alexandra Carpentier[3], Alessandro Lazaric[4], Michal Valko[2]

[1]Lelivrescolaire.fr
[2]SequeL team, INRIA Lille - Nord Europe
[3]Otto-von-Guericke-Universität Magdeburg
[4]Facebook Artifical Intelligence Research

Abstract

In bandits, arms' distributions are stationary. This is often violated in practice, where rewards change over time. In applications as recommendation systems, online advertising, and crowdsourcing, the changes may be triggered by the pulls, so that the arms' rewards change as a function of the number of pulls. In this paper, we consider the specific case of *non-parametric rotting bandits*, where the expected reward of an arm may decrease every time it is pulled. We introduce the *filtering on expanding window average* (`FEWA`) algorithm that at each round constructs moving averages of increasing windows to identify arms that are more likely to return high rewards when pulled once more. We prove that, without any knowledge on the decreasing behavior of the arms, `FEWA` achieves similar anytime problem-dependent, $\widetilde{\mathcal{O}}(\log(KT))$, and problem-independent, $\widetilde{\mathcal{O}}(\sqrt{KT})$, regret bounds of near-optimal stochastic algorithms as `UCB1` of Auer et al. (2002a). This result substantially improves the prior result of Levine et al. (2017) which needed knowledge of the horizon and decaying parameters to achieve problem-independent bound of only $\widetilde{\mathcal{O}}(K^{1/3}T^{2/3})$. Finally, we report simulations confirming the theoretical improvements of `FEWA`.

1 Introduction

Multi-arm bandits (Thompson, 1933; Cesa-Bianchi and Lugosi, 2006; Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2019) formalizes the core aspects of the exploration-exploitation dilemma in online learning, where an agent has to trade off the *exploration* of the environment to gather information and the *exploitation* of the current knowledge to maximize the reward. In the *stochastic setting* (Thompson, 1933; Auer et al., 2002a), each arm is characterized by a stationary reward distribution and whenever an agent pulls an arm, it observes an i.i.d. sample from the corresponding distribution. Despite the extensive algorithmic and theoretical study of this setting (Cesa-Bianchi and Lugosi, 2006; Bubeck and Cesa-Bianchi, 2012; Kaufmann et al., 2012; Garivier and Cappé, 2011), the stationarity assumption is often too restrictive in practice, since the value of the arms may change over time (e.g., change of the preferences of users). The *adversarial setting* (Auer et al., 2002b) addresses this limitation by removing any assumption on how the rewards are generated and learning agents should be able to perform well for any *arbitrary* sequence of rewards. While algorithms such as Exp3 (Auer et al., 2002b) are guaranteed to achieve small regret in this setting, their behavior is conservative as all arms are repeatedly explored in order to avoid incurring too much regret because of unexpected changes in arms' values, which corresponds to unsatisfactory performance in practice, where arms values, while non-stationary, are far from being adversarial. Garivier and Moulines (2011) proposed a variation of the stochastic setting, where the distribution of each arm is *piecewise stationary*. Similarly, Besbes et al. (2014) introduced an adversarial setting where the total amount of change in arms' values is bounded. While these settings effectively capture the characteristics of a wide set of applications, they consider the case where the arms' value evolves *independently* from the decisions of the agent. This setting is often called *restless* bandits. On the other hand, in many problems, the value of an arm changes only when it is pulled and we talk about *rested* bandits. For instance, the value of a service may deteriorate only when it is actually used. Next, if a recommender system shows always the same item to the users, they get bored and enjoy less their experience on the platform. Finally, a student can master a frequently taught topic in an intelligent tutoring system and extra learning on that topic would be less effective. A particularly interesting case is represented by the *rotting bandits*, where the value of an arm decreases every time it is pulled. More precisely, each reward is non-increasing, since it could remain constant at each pull. Heidari et al. (2016)

studied this problem in the case where the rewards observed by the agent are deterministic (i.e., no noise) and showed how a greedy policy (i.e., selecting the arm that returned the largest reward the last time it was pulled) is optimal up to a small constant factor depending on the number of arms K and the largest per-round decay in the arms' value L. Bouneffouf and Féraud (2016) considered the stochastic setting when the dynamics of the rewards is known up to a constant factor. Finally, Levine et al. (2017) defined both non-parametric and parametric noisy rotting bandits, for which they derive new algorithms with regret guarantees. In particular, in the non-parametric case, where the decrease in reward is neither constrained nor known, they introduce the *sliding-window average* (`wSWA`) algorithm, which is shown to achieve a regret to the optimal policy of order $\widetilde{\mathcal{O}}(K^{1/3}T^{2/3})$, where T is the number of rounds in the experiment.

In this paper, we study the non-parametric rotting setting of Levine et al. (2017) and introduce *Filtering on Expanding Window Average* (`FEWA`) algorithm, a novel method that at each round constructs moving average estimates with different windows to identify the arms that are more likely to perform well if pulled once more. Under the assumption that the reward decay are bounded by L, we show that `FEWA` achieves a regret of $\widetilde{\mathcal{O}}(\sqrt{KT})$ without any prior knowledge of L, thus *significantly improving over* `wSWA` and matching the minimax rate of stochastic bandits up to logarithmic factor. This shows that learning with non-increasing rewards is not more difficult than in the constant case (the stochastic setting). Furthermore, when rewards are constant *we recover standard problem-dependent* `UCB` *regret guarantees* (up to constants), while in the rotting bandit scenario with no noise, the regret reduces to the one derived by Heidari et al. (2016). Finally, numerical simulations confirm our theoretical result and show the superiority of `FEWA` over `wSWA`.

2 Preliminaries

We consider a rotting bandits similar to the ones introduced by Levine et al. (2017). At each round t, an agent chooses an arm $i(t) \in \mathcal{K} = \{1, ..., K\}$ and it receives a noisy reward $r_{i(t),t}$. Unlike in standard bandits, the reward associated to each arm i is a σ^2-sub-Gaussian random variable with an expected value $\mu_i(n)$, which depends on the number of times n it was pulled before, e.g., $\mu_i(0)$ is the expectation at the beginning.[1] More formally, let $\mathcal{H}_t \triangleq \big\{\big\{i(s), r_{i(s),s}\big\}, \forall s < t\big\}$ be the sequence of arms pulled and reward observed over time until round t ($\mathcal{H}_0 = \emptyset$), then

$$r_{i(t),t} \triangleq \mu_{i(t)}(N_{i(t),t}) + \varepsilon_t \quad \text{with } \mathbb{E}[\varepsilon_t|\mathcal{H}_t] = 0$$
$$\text{and } \forall \lambda \in \mathbb{R}, \quad \mathbb{E}\big[e^{\lambda \varepsilon_t}\big] \leq e^{\frac{\sigma\lambda^2}{2}},$$

where $N_{i,t} = \sum_{s=1}^{t-1} \mathbb{I}\{i(t) = i\}$ is the number of times arm i is pulled before round t. In the following, by $r_i(n)$ we also denote the random reward obtained from arm i when it is pulled for the n-th time, e.g., $r_{i(t),t} = r_{i(t)}(N_{i(t),t})$. We finally introduce a non-parametric rotting assumption with bounded decay.

Assumption 1. *The reward functions μ_i are non-increasing with bounded decays $-L \leq \mu_i(n+1) - \mu_i(n) \leq 0$. For the sake of the analysis, we also assume that the first pull is bounded $\forall i \in [K]$ $\mu_i(0) \in [0, L]$. We refer to this set of functions as $\mathcal{L}_L$.*

Similarly to Levine et al. (2017), we consider non-increasing functions $\mu_i(n)$ where the value of arms can only decrease when they are pulled. However, we do not restrict them to stay positive but we bound the per-round decay by L. On one hand, any function in $\mathcal{L}_L$ has the range bounded in $[-LT, L]$. Therefore, our setting is included in the setting of Levine et al. (2017) when $\mu_{\max} \triangleq L(T+1)$. However, the regret of `wSWA`, defined below in Equation 2, is bounded by $\widetilde{\mathcal{O}}(\mu_{\max}^{1/3} K^{1/3} T^{2/3})$ which becomes $\mathcal{O}(T)$ in our setting. Therefore, `wSWA` is not proved to learn in our setting. On the other hand, any decreasing function with range in $[0, \mu_{\max}]$ is included in $\mathcal{L}_L$ for $L \triangleq \mu_{\max}$. Therefore, our analysis applies directly to the setting of Levine et al. (2017) by simply setting $L \triangleq \mu_{\max}$, where we get a regret bound of $\widetilde{\mathcal{O}}(\sqrt{KT})$ thereby significantly improving the rate on their result.

The learning problem In general, an agent's policy π returns the arm to pull at round t on the basis of the whole history of observations, i.e., $\pi(\mathcal{H}_t) \in \mathcal{K}$. In the following, we use $\pi(t)$ as shorthand notation

[1] Our definition of $\mu_i(n)$ slightly differs from Levine et al. (2017), where it denotes the expected value of arm i when it is pulled *for the n-th time* instead of *after n pulls*. As a result, in Levine et al. (2017), define $\mu_i(n)$ from $n = 1$, while with our notation it actually starts from $n = 0$.

for $\pi(\mathcal{H}_t)$. The performance of a policy π is measured by the (expected) rewards accumulated over time,

$$J(T,\pi) \triangleq \sum_{t=1}^{T} \mu_{\pi(t)}\big(N_{\pi(t),t}\big).$$

Since π depends on the (random) history observed over time, $J(T,\pi)$ is also random. We therefore define the expected cumulative reward as $\overline{J}(T,\pi) = \mathbb{E}\big[J(T,\pi)\big]$. We restate a useful characterization of the optimal policy given by Heidari et al. (2016).

Proposition 1. *If the (exact) mean of each arm is known in advance for any number of pulls, then the optimal policy $\pi^\star$ maximizing the expected cumulative reward $\overline{J}(T,\pi)$ is greedy at each round, i.e.,*

$$\pi^\star(t) = \arg\max_i \mu_i(N_{i,t}). \tag{1}$$

We denote by $J^\star = \overline{J}(T,\pi^\star) = J(T,\pi^\star)$, the cumulative reward of the optimal policy.

The objective of a learning algorithm is to implement a policy π whose performance is close to $\pi^\star$'s as much as possible. We define the (random) regret as

$$R_T(\pi) \triangleq J^\star - J(\pi,T). \tag{2}$$

Notice that the regret is measured against an optimal allocation over arms rather than a fixed-arm policy as it is a case in adversarial and stochastic bandits. Therefore, even the adversarial algorithms that one could think of applying in our setting (e.g., `Exp3` of Auer et al., 2002a) are not known to provide any guarantee for our definition of regret. On the other hand, for constant $\mu_i(n)$ our problem reduces to standard stochastic bandits. Therefore, our regret definition reduces to the standard stochastic regret. Therefore, for constant functions, any algorithm with some guarantee for rotting regret immediately inherits the same guarantee for the standard regret.

Let $N^\star_{i,T}$ be the (deterministic) number of times that arm i is pulled by the optimal policy $\pi^\star$ up to time T (excluded). Similarly, for a given policy π, let $N^\pi_{i,T}$ be the (random) number pulls of arm i. Using this notation, notice that the cumulative reward can be rewritten as

$$J(T,\pi) = \sum_{t=1}^{T}\sum_{i\in\mathcal{K}} \mathbb{I}_{\{\pi(t)=i\}}\mu_i\big(N^\pi_{i,t}\big) = \sum_{i\in\mathcal{K}}\sum_{s=0}^{N^\pi_{i,T}} \mu_i(s).$$

Then, we can conveniently rewrite the regret as

$$R_T(\pi) = \sum_{i\in\mathcal{K}}\left(\sum_{s=0}^{N^\star_{i,T}}\mu_i(s) - \sum_{s=0}^{N^\pi_{i,T}}\mu_i(s)\right) = \sum_{i\in \text{UP}}\sum_{s=N^\pi_{i,T}+1}^{N^\star_{i,T}}\mu_i(s) - \sum_{i\in \text{OP}}\sum_{s=N^\star_{i,T}+1}^{N^\pi_{i,T}}\mu_i(s), \tag{3}$$

where $\text{UP} = \big\{i\in\mathcal{K} | N^{\pi^\star}_{i,T} > N^\pi_{i,T}\big\}$ and $\text{OP} = \big\{i\in\mathcal{K} | N^{\pi^\star}_{i,T} < N^\pi_{i,T}\big\}$ are the sets of arms that are respectively under-pulled and over-pulled by π w.r.t. the optimal policy.

Prior regret bounds In order to ease the discussion of the theoretical results we derive in Sect. 4, we restate prior results for two special cases. We start with the minimax regret lower bound for stochastic bandits, which corresponds to the case when the expected rewards $\mu_i(n)$ are constant.

Proposition 2. *(Auer et al., 2002b, Thm. 5.1) For any learning policy π and any horizon T, there exists a stochastic stationary problem $\{\mu_i(n) = \mu_i\}_i$ with K sub-Gaussian arms with parameter σ such that π suffers an expected regret*

$$\mathbb{E}[R_T(\pi)] \geq \frac{\sigma}{10}\min\Big(\sqrt{KT}, T\Big).$$

where the expectation is taken with respect to both the randomization over rewards and the algorithms internal randomization,

Proposition 2 can also be proved without the randomization device. The constant $1/10$ in the lower bound above can be improved to $1/4$ (Cesa-Bianchi and Lugosi, 2006, Theorem 6.11).

Next, Heidari et al. (2016) previously derived lower and upper bounds for the regret in the case of deterministic rotting bandits (i.e., $\sigma = 0$).

Proposition 3. *(Heidari et al., 2016, Thm. 3) For any learning policy π, there exists a deterministic rotting bandits (i.e., $\sigma = 0$) satisfying Assumption 1 with bounded decay L such that π suffers an expected regret*

$$\mathbb{E}[R_T(\pi)] \geq \frac{L}{2}(K-1).$$

Let π^{σ_0} be a greedy (not necessarily an oracle) policy that selects at each round the arm with the largest upcoming reward $\arg\max_i(\mu_i(N_{i,t}-1))$. For any deterministic rotting bandits (i.e., $\sigma = 0$) satisfying Assumption 1 with bounded decay L, π^{σ_0} suffers an expected regret

$$\mathbb{E}[R_T(\pi^{\sigma_0})] \leq L(K-1).$$

Propositions 2 and 3 bound the performance of any algorithm on the constant and deterministic classes of problems with respective parameters σ and L. Note that any problem in one of these two classes is a rotting problem with parameters (σ, L). Therefore, the performance of any algorithm on the rotting problem described above is also bounded by both lower bounds.

3 FEWA: Filtering on Expanding Window Average

Since the expected rewards μ_i change over time, the main difficulty in the non-parametric rotting bandit setting introduced in the previous section is that we cannot entirely rely on all the samples observed until time t to accurately predict which arm is likely to return the highest reward in the future. In particular, the older the sample, the less representative is of the reward that the agent may observe by pulling the same arm once again. This suggests that we should construct estimates using the more recent samples. On the other hand, by discarding older rewards, we also reduce the number of samples used in the estimates, thus increasing their variance. In Algorithm 1 we introduce a novel algorithm (`FEWA` or π_{F}) that at each round t, relies on estimates using windows of increasing length to filter out arms that are suboptimal with high probability and then pulls the least pulled arm among the remaining arms.

Before we describe `FEWA` in detail, we first describe the subroutine `Filter` in Algorithm 2, which receives as input a set of active arms $\mathcal{K}_h$, a window h, and a confidence parameter δ, to return an updated set of arm $\mathcal{K}_{h+1}$. For each arm i that has been pulled n times, the algorithm constructs an estimate $\widehat{\mu}_i^h(n)$ that averages the h most recent rewards observed from i. The estimator is well defined only for $h \leq n$. Nonetheless, the construction of the set $\mathcal{K}_h$ and the stopping condition at Line 10 in Algorithm 1 guarantee that $\widehat{\mu}_i^h(N_{i,t})$ are always well defined for the arms in $\mathcal{K}_h$. The subroutine `Filter` then discards from $\mathcal{K}_h$ all the arms whose mean estimate (built with window h) is lower than the empirically best arm by more than twice a threshold $c(h, \delta_t)$ constructed by standard Hoeffding's concentration inequality (see Algorithm 4).

Algorithm 1 `FEWA`

Input: σ, $\mathcal{K}$, δ_0, α

1: pull each arm once, collect reward, and initialize $N_{i,K} \leftarrow 1$
2: **for** $t \leftarrow K+1, K+2, \ldots$ **do**
3: $\quad\delta_t \leftarrow \delta_0/(Kt^\alpha)$
4: $\quad h \leftarrow 1$ {*initialize bandwidth*}
5: $\quad\mathcal{K}_1 \leftarrow \mathcal{K}$ {*initialize with all the arms*}
6: $\quad i(t) \leftarrow$ `none`
7: $\quad$**while** $i(t)$ is `none` **do**
8: $\quad\quad\mathcal{K}_{h+1} \leftarrow$ `Filter`$(\mathcal{K}_h, h, \delta_t)$
9: $\quad\quad h \leftarrow h+1$
10: $\quad\quad$**if** $\exists i \in \mathcal{K}_h$ such that $N_{i,t} = h$ **then**
11: $\quad\quad\quad i(t) \leftarrow i$
12: $\quad\quad$**end if**
13: $\quad$**end while**
14: $\quad$receive $r_i(N_{i,t+1}) \leftarrow r_{i(t),t}$
15: $\quad N_{i(t),t} \leftarrow N_{i(t),t-1} + 1$
16: $\quad N_{j,t} \leftarrow N_{j,t-1}, \quad \forall j \neq i(t)$
17: **end for**

Algorithm 2 `Filter`

Input: $\mathcal{K}_h$, h, δ_t

1: $c(h, \sigma, \delta_t) \leftarrow \sqrt{(2\sigma^2/h)\log(1/\delta_t)}$
2: **for** $i \in \mathcal{K}_h$ **do**
3: $\quad \widehat{\mu}_i^h(N_{i,t}) \leftarrow \frac{1}{h}\sum_{j=1}^{h} r_i(N_{i,t} - j)$
4: **end for**
5: $\widehat{\mu}_{\max,t}^h \leftarrow \max_{i \in \mathcal{K}_h} \widehat{\mu}_i^h(N_{i,t})$
6: **for** $i \in \mathcal{K}_h$ **do**
7: $\quad \Delta_i \leftarrow \widehat{\mu}_{\max,t}^h - \widehat{\mu}_i^h(N_{i,t})$
8: $\quad$ **if** $\Delta_i \leq 2c(h, \sigma, \delta_t)$ **then**
9: $\quad\quad$ add i to $\mathcal{K}_{h+1}$
10: $\quad$ **end if**
11: **end for**

Output: $\mathcal{K}_{h+1}$

The `Filter` subroutine is used in `FEWA` to incrementally refine the set of active arms, starting with a window of size 1, until the condition at Line 10 is met. As a result, $\mathcal{K}_{h+1}$ only contains arms that passed the filter for all windows from 1 up to h. Notice that it is crucial to start filtering arms from a small window and to keep refining the previous set of active arms, instead of completely recomputing them for every new window h. In fact, the estimates constructed using a small window use recent rewards, which are closer to the future value of an arm. As a result, if there is enough evidence that an arm is suboptimal already at a small window h, then there is no reason to consider it again for larger windows. On the other hand, a suboptimal arm may pass the filter for small windows as the threshold $c(h, \sigma, \delta_t)$ is large for small h, i.e., when only a few samples are used in constructing $\widehat{\mu}_i^h(N_{i,t})$. Thus, `FEWA` keeps refining $\mathcal{K}_h$ for larger and larger windows in the attempt of constructing more and more accurate estimates and discard more suboptimal arms. This process stops when we reach a window as large as the number of samples for at least one arm in the active set $\mathcal{K}_h$ (i.e., Line 10). At this point, increasing h would not bring any additional evidence that could refine $\mathcal{K}_h$ further[2] and `FEWA` finally selects the active arm $i(t)$ whose number of samples matches the current window, i.e., the least pulled arm in $\mathcal{K}_h$. The set of available rewards and the number of pulls are then updated accordingly.

4 Analysis

We first state the major theoretical result of the paper, the problem-independent bound for `FEWA` and then sketch the proof in Section 4.1. Then, in Section 4.2, we give problem-dependent guarantees.

Theorem 1. *For any rotting bandit scenario with means* $\{\mu_i(n)\}_{i,n}$ *satisfying Assumption 1 with bounded decay* L *and any time horizon* T, `FEWA` *run with* $\alpha = 5$, $\delta_0 = 1$, *i.e., with* $\delta_t = 1/(Kt^5)$, *suffers an expected regret*[3]

$$\mathbb{E}[R_T(\pi_F)] \leq 13\sigma(\sqrt{KT} + K)\sqrt{\log(KT)} + KL.$$

Theorem 1 shows that `FEWA` achieves a $\widetilde{\mathcal{O}}(\sqrt{KT})$ regret *without any knowledge* of the size of decay L. This significantly improves over the regret of `wSWA` (Levine et al., 2017), which is of order $\widetilde{\mathcal{O}}\big(K^{1/3}T^{2/3}\big)$ and *needs to know* L. The improvement is also due to the fact that `FEWA` exploits filters using moving averages with increasing windows to discard arms that are with high probability suboptimal. Since this process is done at each round, `FEWA` smoothly tracks changes in the value of each arm, so that if an arm becomes worse later on, other arms would be recovered and pulled again. On the other hand, `wSWA` relies on a fixed exploratory phase where all arms are pulled in a round-robin fashion and the tracking is performed using averages constructed with a fixed window. Furthermore, while the performance of `wSWA` can be optimized by having prior knowledge on the range of the expected rewards (see the tuning of α in the work of Levine et al. 2017, Theorem 3.1), `FEWA` does not require any knowledge of L to achieve the $\widetilde{\mathcal{O}}(\sqrt{KT})$ regret. Moreover, `FEWA` in naturally anytime (T does not need to be known), while the fixed exploratory phase of `wSWA` requires T to be properly tuned and resorts to a doubling trick to be anytime. Algorithms (such as `FEWA`) with direct anytime guarantees show a practical advantage over the doubling-trick ones, that often give a suboptimal empirical performance.

[2] $\widehat{\mu}_i^h(N_{i,t})$ is not defined for $h > N_{i,t}$
[3] See Corollary 3 and 4 for the high-probability result.

For $\sigma = 0$, our upper bound reduces to KL, thus matching the prior (upper and lower) bound of Heidari et al. (2016) for deterministic rotting bandits. Moreover, the additive decomposition of regret shows that there is *no coupling* between the stochastic problem and the rotting problem as the σ terms are summed with the L term while `wSWA` shows an $L^{1/3}\sigma^{2/3}$ factor[4] in front of the leading term. Finally, the $\mathcal{O}(\sqrt{KT\log T})$ matches the worst-case optimal regret bound of the standard stochastic bandits (i.e., $\mu_i(n)$s are constant) up to a logarithmic factor. Whether an algorithm can achieve $\mathcal{O}(\sqrt{KT})$ regret bound is an open question. On one hand, `FEWA` uses more confidence bounds than `UCB1` to track change for each arm. Thus, `FEWA` uses larger bands in order to make all the confidence bounds hold with high probability. Therefore, we pay an extra exploration cost which may be necessary for handling the possible rotting behavior of arms. On the other hand, our worst-case analysis shows that some of the difficult problems that reach the worst-case bound of Theorem 1 are realized with constant functions, which is the standard stochastic bandits. For standard stochastic bandits, it is known that `MOSS`-like (Audibert and Bubeck, 2009) strategies are able to get regret guarantees without the $\log T$ factor. To sum up, the necessity of the extra $\log T$ factor for the worst-case regret of rotting bandits remains an open problem.

4.1 Sketch of the proof

In this section, we give a sketch of the proof of the regret bound. We first introduce the expected value of the estimators used in `FEWA`. For any n and $h \leq n$, we define

$$\overline{\mu}_i^h(n) \triangleq \mathbb{E}\big[\widehat{\mu}_i^h(n)\big] = \frac{1}{h}\sum_{j=1}^{h}\mu_i(n-j).$$

Notice that if at round t, the number of pulls to arm i is $N_{i,t}$, then $\overline{\mu}_i^1(N_{i,t}) = \mu_i(N_{i,t}-1)$, which is the expected value of arm i the last time it was pulled. We now use Hoeffding's concentration inequality and the favorable events that we consider throughout the analysis.

Proposition 4. *For any fixed arm i, number of pulls n and window h, we have with probability $1-\delta$,*

$$\left|\widehat{\mu}_i^h(n) - \overline{\mu}_i^h(n)\right| \leq c(h,\delta) \triangleq \sqrt{\frac{2\sigma^2}{h}\log\frac{1}{\delta}}. \tag{4}$$

Furthermore, for any round t, for a confidence $\delta_t \triangleq \delta_0/(Kt^\alpha)$, let

$$\xi_t \triangleq \left\{\forall i \in \mathcal{K}, \forall n \leq t, \forall h \leq n, \left|\widehat{\mu}_i^h(n) - \overline{\mu}_i^h(n)\right| \leq c(h,\delta_t)\right\}$$

be the event under which all the possible estimates constructed by `FEWA` at round t are well concentrated towards their expected value. Then, taking the union bound, $\mathbb{P}(\xi_t) \geq 1 - Kt^2\delta_t/2$.

Quality of arms in the active set We are now ready to derive a crucial lemma that provides support to the arm selection process implemented by `FEWA` through the series of refinements obtained by the `Filter` subroutine. Recall that at any round t, after pulling arms $\{N_{i,t}^{\pi_F}\}_i$ the greedy (oracle) policy would select an arm characterized by

$$i_t^\star\left(\left\{N_{i,t}^{\pi_F}\right\}_i\right) \in \arg\max_{i\in\mathcal{K}} \mu_i\left(N_{i,t}^{\pi_F}\right).$$

We denote by $\mu_t^+(\pi_F) \triangleq \max_{i\in\mathcal{K}}\mu_i(N_{i,t}^{\pi_F})$, the expected reward that such oracle policy would obtain by pulling $i_t^\star$. Notice that the dependence on π_F in the definition of $\mu_t^+(\pi_F)$ is due to the fact that we consider what the deterministic oracle policy would do at the state reached by π_F. While `FEWA` cannot directly target the performance of the greedy arm, the following lemma shows that the last h pulls of any arms in the active set returned by the filter are close to the performance of the current best arm up to four times the confidence band $c(h,\delta_t)$.

Lemma 1. *On favorable event ξ_t, if an arm i passes through a filter of window h at round t, the average of its h last pulls cannot deviate significantly from the best available arm $i_t^\star$ at that round, i.e.,*

$$\overline{\mu}_i^h(N_{i,t}) \geq \mu_t^+(\pi_F) - 4c(h,\delta_t).$$

[4] Specifically, it is $\mu_{\max}^{1/3}\sigma^{2/3}$, where $\mu_{\max}$ is equivalent to L in our setting, though our setting is more general as explained in the remark following Assumption 1.

Relating FEWA to the optimal policy While Lemma 1 (with proof in the appendix) provides a first link between the value of the arms returned by the filter and the greedy arm, $i_t^\star$ is still defined according to the number of pulls obtained by FEWA up to t. On the other hand, the optimal policy could actually pull a different sequence of arms and at t it could have different number of pulls. In order to bound the regret, we need to relate the actual performance of the optimal policy to the value of the arms pulled by FEWA. We let $h_{i,t} \triangleq \left|N_{i,t}^{\pi_\mathrm{F}} - N_{i,t}^{\pi^\star}\right|$ be the absolute difference in the numbers of pulls between π_F and the optimal policy. Since $\sum_{i\in\mathrm{OP}} N_{i,t}^{\pi_\mathrm{F}} = \sum_{i\in\mathrm{UP}} N_{i,t}^{\pi^\star} = t$, we have that $\sum_{i\in\mathrm{OP}} h_{i,t} = \sum_{i\in\mathrm{UP}} h_{i,t}$ which means that there are as many overpulls than underpulls over all arms. Let $j \in \mathrm{UP}$ be an underpulled arm[5] with $N_{j,T}^{\pi_\mathrm{F}} < N_{j,T}^{\pi^\star}$. Then, we have the inequalities

$$\forall s \in \{1,\dots,h_{i,t}\},\ \mu_T^+(\pi_\mathrm{F}) = \max_{i\in\mathcal{K}} \mu_i(N_{i,T}^{\pi_\mathrm{F}}) \geq \mu_j(N_{j,T}^{\pi_\mathrm{F}} + s). \tag{5}$$

As a consequence, we derive the first upper bound on the regret from Equation 3 as

$$R_T(\pi_\mathrm{F}) = \sum_{i\in\mathrm{UP}} \sum_{t'=N_{i,T}^{\pi_\mathrm{F}}+1}^{N_{i,T}^{\pi^\star}} \mu_i(t') - \sum_{i\in\mathrm{OP}} \sum_{t'=N_{i,T}^{\pi^\star}+1}^{N_{i,T}^{\pi_\mathrm{F}}} \mu_i(t') \leq \sum_{i\in\mathrm{OP}} \sum_{h=0}^{h_{i,T}-1} \Big(\mu^+(\pi_\mathrm{F}) - \mu_i(N_{i,T}^{\pi^\star}+h)\Big), \tag{6}$$

where the inequality is obtained by bounding $\mu_i(t') \leq \mu_T^+(\pi_\mathrm{F})$ in the first summation[6] and then using $\sum_{i\in\mathrm{OP}} h_{i,T} = \sum_{i\in\mathrm{UP}} h_{i,T}$. While the previous expression shows that we can now only focus on over-pulled arms in OP, it is still difficult to directly control the expected reward $\mu_i(N_{i,T}^{\pi^\star}+h)$, as it may change at each round (by at most L). Nonetheless, we notice that its cumulative sum can be directly linked to the average of the expected reward over a suitable window. In fact, for any $i \in \mathrm{OP}$ and $h_{i,T} \geq 2$, we have

$$(h_{i,T}-1)\overline{\mu}_i^{h_{i,T}-1}(N_{i,T}-1) = \sum_{t'=0}^{h_{i,T}-2} \mu_i(N_{i,T}^{\pi^\star}+t').$$

At this point we can control the regret for each $i \in \mathrm{OP}$ in Equation 6 by applying the following corollary derived from Lemma 1.

Corollary 1. *Let $i \in \mathrm{OP}$ be an arm overpulled by FEWA at round t and $h_{i,t} \triangleq N_{i,t}^{\pi_\mathrm{F}} - N_{i,t}^{\pi^\star} \geq 1$ be the difference in the number of pulls w.r.t. the optimal policy $\pi^\star$ at round t. On favorable event ξ_t, we have*

$$\mu_t^+(\pi_\mathrm{F}) - \overline{\mu}_i^{h_{i,t}}(N_{i,t}) \leq 4c(h_{i,t},\sigma,\delta_t). \tag{7}$$

4.2 Discussion on problem-dependent result and the price of decaying rewards

Since our setting generalizes the standard bandit setting, where $\{\mu_i\}_i$ are constant over pulls, a natural question is whether we pay any price for this generalization. While the result of Levine et al. (2017) suggested that learning in rotting bandits could be more difficult, in Theorem 1, we proved that FEWA matches the minimax regret $\widetilde{\mathcal{O}}(\sqrt{KT})$ for multi-arm bandits.

However, we may now wonder whether FEWA also matches the result of, e.g., UCB in terms of *problem-dependent* regret. As illustrated in the next remark, we show that up to constants, FEWA performs as well as UCB on any stochastic problem.

Remark 1. *If we apply the result of Corollary 1 applied to stochastic bandits, i.e., when μ_i are constant and $\mu_\star \triangleq \max_i \mu_i$, we get that for $\delta_t \geq 1/(KT^\alpha)$,*

$$\mu_\star - \mu_i \leq 4c(h_{i,T}-1,\delta_t) = 4\sqrt{\frac{2\alpha\sigma^2\log(KT)}{h_{i,T}-1}} \text{ or equivalently, } h_{i,T} \leq 1 + \frac{32\alpha\sigma^2\log(KT)}{(\mu_\star-\mu_i)^2}. \tag{8}$$

Therefore, our algorithm matches the lower bound of Lai and Robbins (1985) up to a constant. Moreover, in the case of constant functions, our upper bound for FEWA is at most α larger than the one for UCB1 (Auer

[5] if such arm does not exist, then π_F suffers no regret

[6] notice that since $t' \geq N_{i,T}^{\pi_\mathrm{F}} + 1$ and μ_i is decreasing, the inequality directly follows from the definition of $\mu_T^+(\pi_\mathrm{F})$

et al., 2002a).[7] The main source of suboptimality is the use of a confidence bound filtering instead of an upper-confidence index policy. Selecting the less pulled arm in the active set is conservative as it requires uniform exploration until elimination, resulting in factor 4 in the confidence bound guarantee on the selected arm (versus 2 for `UCB`) which implies 4 times more overpulls than `UCB` (see Equation 8). We conjecture this may not be necessarily needed and it is an open question whether it is possible to derive either an index policy or a selection rule that is better than pulling the less pulled arm in the active set. The other source of suboptimality w.r.t. `UCB` is the use of larger confidence band because (1) the higher number of estimators computed at each round and (Kt^2 instead of Kt for `UCB`) and because (2) the regret at each round in the worst case grows as Lt, which requires reducing the probability of the unfavorable event.

As a result of Remark 1, we claim, that surprisingly and contrarily to what the prior work (Levine et al., 2017) suggests, the rotting bandits are *not significantly more difficult* than the multi-arm bandits with constants mean rewards. We show this observation is not only theoretical. In particular, in Section 5, we show that in our experiments, the empirical regret of `FEWA` was at most twice as large as `UCB1`.

Remark 1 also reveals that Corollary 1 is in fact a problem-dependent result. Similarly, as we derived a problem-dependent bound of `FEWA`'s regret for constant functions (standard stochastic bandits) we now show a way to get a similar problem-dependent bound for the general case. In particular, with Corollary 1 we upper-bound the maximum number of overpulls by a problem dependent quantity

$$h^+_{i,T} \triangleq \max\left\{h \leq 1 + \frac{32\alpha\sigma^2\log(KT)}{\Delta^2_{i,h-1}}\right\}, \text{ where } \Delta_{i,h} \triangleq \min_{j\in\mathcal{K}} \mu_j\left(N^\star_{j,T}-1\right) - \overline{\mu}^h_i\left(N^\star_{i,t}+h\right). \tag{9}$$

We then use Corollary 1 again to upper-bound the regret caused by $h^+_{i,T}$ overpulls for each arm, leading to Corollary 2. The complete proof is in Appendix D.

Corollary 2 (problem-dependent guarantee)**.** *For $\delta_t \triangleq 1/(Kt^5)$, the regret is bounded as*

$$\mathbb{E}[R_T(\pi_F)] \leq \sum_{i\in\mathcal{K}}\left(\frac{C_5\log(KT)}{\Delta_{i,h^+_{i,T}-1}} + \sqrt{C_5\log(KT)} + L\right) \text{ with } C_\alpha \triangleq 32\alpha\sigma^2 \text{ and } h^+_{i,T} \text{ defined in Equation 9.}$$

4.3 Runtime and memory usage

At each round t, `FEWA` has a worst-case time and memory complexity of a $\mathcal{O}(t)$. In fact, it needs to store and update up to t averages per-arm. Since moving from an average computed on window h to $h+1$ can be done at a cost $\mathcal{O}(1)$ the per-round complexity is $\mathcal{O}(T)$. Such complexity may be undesirable.[8]

The first idea to improve time and memory complexity is to reduce the number of filters used in the selection. We first notice that the selectivity of the filters scales with $1/\sqrt{h}$. As a result, when h increases, the usefulness of the consecutive filters decreases. This remark suggests that we could replace the window increment (Line 9 of Algorithm 1) by a geometric update with factor 2 for time t in order to have a constant ratio between two selectivity values. However, this is not enough to reduce the amount of computation. In fact, we still have to compute ($\log_2 T$ number of) averages of up T samples and therefore we still pay $\mathcal{O}(T)$ in time and memory. We therefore provide a more efficient version of `FEWA`, called `EFF-FEWA` (Appendix E) which also uses $\log_2 T$ filters (handling the expanding dynamics) but now with precomputed statistics (handling the sliding dynamics) only being updated when the number of samples for a particular arm *doubles*. Specifically, the precomputed statistics are updated with a delay in order to be representative of exactly h samples with $h = 2^j$ for some j. For instance, the (two) statistics of length 2 are replaced every 2 pulls while statistics of length 4 are replaced every 4 pulls. Therefore each filter $j \in \{1,\dots,\log_2 T\}$ needs to only store two statistics for each arm $i \in \mathcal{K}$: the *currently* used one $\widehat{s}^c_{i,j}$ and the *pending* one $\widehat{s}^p_{i,j}$. Therefore, at any time, the j-th filter is fed with $\widehat{s}^c_{i,j}$ for all arms i which are averages of 2^{j-1} consecutive samples among the 2^j-1 last ones. In the worst case, the last $2^{j-1}-1$ samples are not covered by filter j but these samples are necessarily covered by all the filters before. This way, `EFF-FEWA` recovers the same bound than `FEWA` up to a constant factor (proof in Appendix E). In contrast, the small number of filters can now be updated sporadically, thus reducing a per-round time and space complexity to only $\mathcal{O}(\log T)$ per arm. A similar yet different idea from the one we propose here has appeared independently in the context of streaming mining (Bifet and Gavaldà, 2007).

[7] To make the results comparable, we need to replace $2\sigma^2$ by $1/2$ in the proof of Auer et al. (2002a) to adapt the confidence bound for a sub-Gaussian noise.

[8] This observation is worst-case. In fact, in some cases, the number of samples for the suboptimal arms may be much smaller than $\mathcal{O}(t)$ For example, in standard bandits it could be $\mathcal{O}(\log t)$. This would dramatically reduce the number of means to compute at each round.

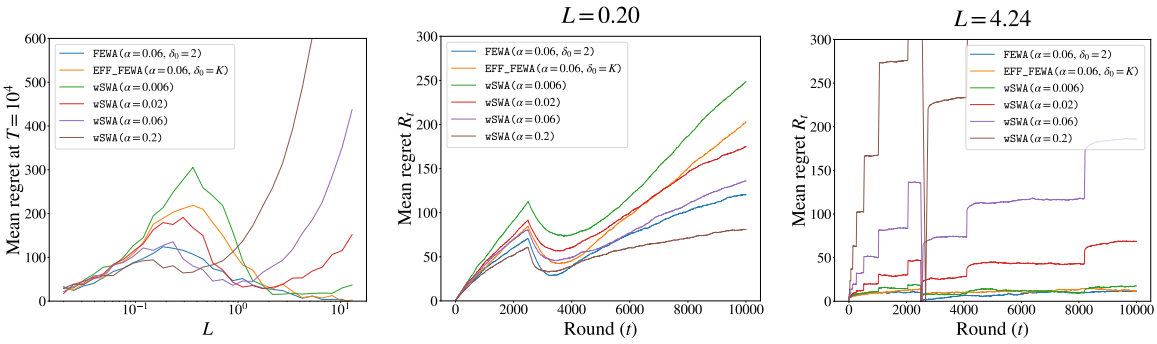

Figure 1: Comparison of the average regret in the two arms single decrement case. **Left:** Regret at the end of the game for a geometric sequence of L. **Middle-right:** Average regret during the game for $L = 0.20$ and $L = 4.24$.

5 Numerical simulations

In this section, we report numerical simulations designed to provide insights on the difference between `wSWA` and `FEWA`. We consider rotting bandits with two arms defined as

$$\mu_1(n) = 0, \quad \forall n \leq T \quad \text{and} \quad \mu_2(n) = \begin{cases} \frac{L}{2} & \text{if } n < \frac{T}{4}, \\ -\frac{L}{2} & \text{if } n \geq \frac{T}{4}. \end{cases}$$

The rewards are then generated by applying a Gaussian i.i.d. noise $\mathcal{N}(0, \sigma = 1)$. The single point of non-stationarity in the second arm is designed to satisfy Figure 1 with a bounded decay L. The gap has been chosen as $T/4$ to not advantage `FEWA`, which pulls each arm $T/2$ times when no arm is filtered. In the two-arms setting defined above, the optimal allocation is $N_{1,T} = 3T/4$ and $N_{2,T} = T/4$.

Both algorithms have a parameter α to tune. In `wSWA`, α is a multiplicative constant for the theoretical optimal window. We try four different values of α, including the recommendation of Levine et al. (2017), $\alpha = 0.2$. In `FEWA`, α tunes the confidence $\delta_t = 1/(t^\alpha)$ of the threshold $c(h, \delta_t)$. While our analysis suggests $\alpha = 5$ (or $\alpha = 4$ for bounded variables), Hoeffding confidence intervals, union bounds, and filtering algorithms are too conservative for a typical case. Therefore, we use a more aggressive $\alpha \triangleq 0.06$. While Theorem 1 suggests that the performance of `FEWA` should only mildly depend on the bounded decay L, Theorem 3.1 of Levine et al. (2017) displays a linear dependence on the largest $\mu_i(0)$, which in this case is L. Their Theorem 3.1 also states that the linear dependence appears for larger L when α is small.

In Figure 1, we validate the difference between the two algorithms and their dependence on L. The first plot shows the regret at the T for various values of L and different algorithms. The second and the third plot shows the regret as a function of the number of rounds for $L = 0.2$ and $L = 4.24$, which correspond to the worst case performance for `FEWA` and to the $L \gg \sigma$ regime. All our experiments are run for $T = 10000$ and we average results over 500 runs.

Before discussing the results, we point out that in the rotting setting, the regret can both increase and decrease over time. Consider two simple policies: π_1, which first pulls arm 1 for $N^\star_{1,T}$ times and the then pulls arm 2 for $N^\star_{2,T}$ times, and π_2 which reverses the order (first arm 2 and then arm 1). If we take π_1 as reference, π_2 would have an increasing regret for the first $T/4$ rounds, which would reverse back to 0 at time $T/2$, since π_2 would select arm 1 getting a reward $L/2$, while π_1 (that had already pulled 1) transitioned to pulling arm 2 with a reward of 0.

As illustrated in Theorem 3.1 of Levine et al. (2017), `wSWA` regret scales linearly with L when $L\alpha \ll 1$. In Figure 1 (left), we show that this regime depends effectively on α: The smaller the α, the smaller the averaging window, the more reactive it is to large drops (see Figure 1, right). On the other hand, `FEWA` ends up doing a single mistake for large L. Therefore, it recovers the $\widetilde{\mathcal{O}}(KL)$ regret with no dependence on T as Heidari et al. (2016). Indeed, when L is large, Corollary 2 shows that, since in our setting, $\Delta_{i,h^+_{i,T}} = L/2$, the leading term is $\widetilde{\mathcal{O}}(KL)$ for a reasonable horizon.

For small L (Figure 1, middle), `wSWA` is competitive only when α is sufficiently large. We see that $\alpha = 0.2$ (recommended by Levine et al., 2017) is indeed a good choice until $L \sim \sigma = 1$, even though it becomes quickly suboptimal after that. For `FEWA`, $L \sim 2\sqrt{K/T}$ corresponds to the hardest problems as suggested by Theorem 1. We conclude that `FEWA` is more robust than `wSWA` as it almost always achieves the best performance across different problems while being agnostic to the value of L. On the other

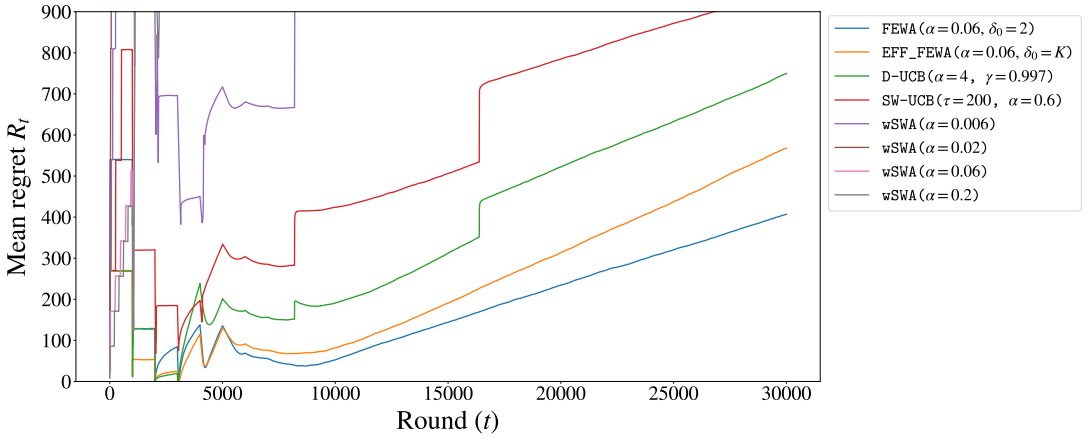

Figure 2: Regret of the setting with 10 arms.

hand, `wSWA`'s performance is very sensitive to the choice of α and the same value of the parameter may correspond to significantly different performance depending on L. Finally, we notice that `EFF-FEWA` has a comparable regret with `FEWA` when L is large, while for a small value of L, `EFF-FEWA` suffers the cost of the delay in its statistics update, which is larger for the last filter.

We also tested our algorithm in a rotting setting with 10 arms: the mean of 1 arm is constant with value 0 while 9 arms after 1000 pulls abruptly decrease from $+\Delta_i$ to $-\Delta_i$. Δ_i is ranging from 0.001 to 10 in a geometric sequence. Figure 2 shows regret for different algorithms. Beside `FEWA` and the four instances of `wSWA`, we add `SW-UCB` and `D-UCB` (Garivier and Moulines, 2011) with window and discount parameters tuned to achieve the best performance. While the two algorithms are known benchmarks for non-stationary bandits, they are designed for the *restless* case. Therefore, they keep exploring arms that have not been pulled for many rounds. This behavior is suboptimal for *rested* bandits that we have here, as the arms stay constant when they are not pulled.

We see that after each switch $+\Delta_i$ to $-\Delta_i$, `FEWA` is among the best ones at quickly recovering and adapting to the new situation. `EFF-FEWA` has similar performance after big drops as it is not too delayed on a new sample. However, the effect of delay in updates has a larger impact in situations where we need many samples to filter an arm. Therefore, we observe a larger regret at the end of the game as compared to `FEWA`. `wSWA` with large α uses windows that are too large and therefore, for very big changes in the mean reward, suffers high empirical regret at the beginning of this game. On the other hand, `wSWA` with small α suffers larger empirical regret at the end of this game where it is blind to small differences between arms, as the window size too small. We conclude that the windows of a fixed size that `wSWA` uses, makes it difficult for `wSWA` to adapt to different situations. Moreover, when α is too large, `wSWA` is very

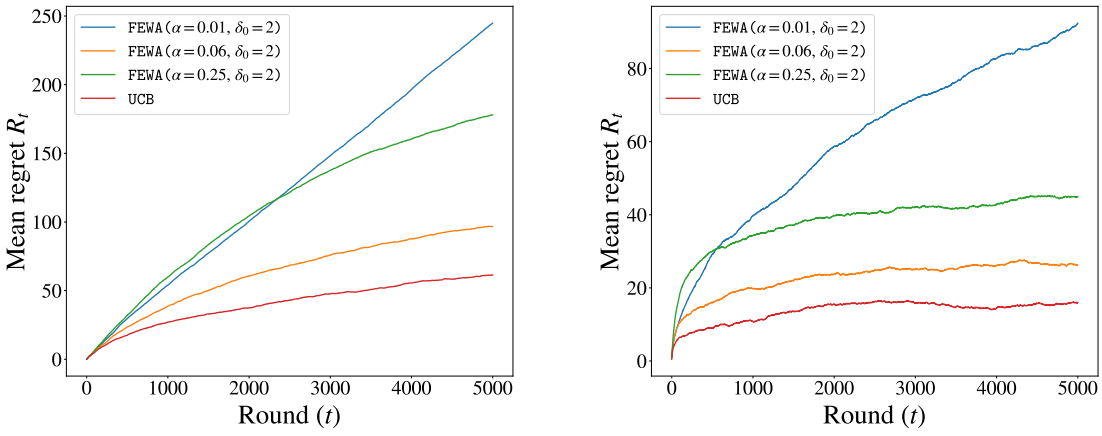

Figure 3: Comparing `UCB1` and `FEWA` with $\Delta = 0.14$ and $\Delta = 1$.

sensitive to its doubling trick.

We remark that SW-UCB and D-UCB show similar behavior. They are both heavily penalized by their restless forgetting even though their forgetting parameters τ and γ are optimally tuned for this experimental setup. Indeed, there is no good choice of parameters as a fast forgetting rate makes the policies repeatedly pull bad arms (whose mean rewards do not change when they are not pulled in our rested setup) while a slow forgetting rate makes the policies not being able to adapt to abrupt shifts.

Finally, in Figure 3 we compare the performance of FEWA against UCB1 (Auer et al., 2002a) on two-arm bandits with different gaps. These experiments confirm the theoretical findings of Theorem 1 and Corollary 2: FEWA has comparable performance with UCB1. In particular, both algorithms have a logarithmic asymptotic behavior and for $\alpha = 0.06$, the ratio between the regret of two algorithms is empirically lower than 2. Notice, the theoretical factor between the two upper bounds is 5 (for $\alpha = 5$). This shows the ability of FEWA to be competitive for stochastic bandits.

6 Conclusion and discussion

We introduced FEWA, a novel algorithm for the non-parametric rotting bandits. We proved that FEWA achieves an $\widetilde{\mathcal{O}}(\sqrt{KT})$ regret without any knowledge of the decays by using moving averages with a window that effectively adapts to the changes in the expected rewards. This result greatly improves the wSWA algorithm proposed by Levine et al. (2017), that suffered a regret of order $\widetilde{\mathcal{O}}(K^{1/3}T^{2/3})$. Our analysis of FEWA is quite non-standard and new. FEWA hinges on the *adaptive* nature of the window size. The most interesting aspect of the proof technique (which can be of independent interest) is that confidence bounds are used not only for the action selection but also for the *data* selection, i.e., to identify the best window to trade off the bias and the variance in estimating the current value of each arm. Furthermore, we show that in the case of constant arms, FEWA recovers the performance of UCB, while in the deterministic case we match the performance of Heidari et al. (2016).

Acknowledgements The research presented was supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, Inria and Otto-von-Guericke-Universität Magdeburg associated-team north-European project Allocate, and French National Research Agency projects ExTra-Learn (n.ANR-14-CE24-0010-01) and BoB (n.ANR-16-CE23-0003). The work of A. Carpentier is also partially supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1), by the DFG - 314838170, GRK 2297 MathCoRe, by the DFG GRK 2433 DAEDALUS, by the DFG CRC 1294 Data Assimilation, Project A03, and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18. This research has also benefited from the support of the FMJH Program PGMO and from the support to this program from CRITEO. Part of the computational experiments was conducted using the Grid'5000 experimental testbed (https://www.grid5000.fr).

References

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, 2009.

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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *Journal on Computing*, 32(1):48–77, 2002b.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed bandit problem with non-stationary rewards. In *Neural Information Processing Systems*, 2014.

Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *International Conference on Data Mining*, 2007.

Djallel Bouneffouf and Raphael Féraud. Multi-armed bandit problem with known trend. *Neurocomputing*, 205(C):16–21, 2016.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5:1–122, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, 2011.

Aurélien Garivier and Eric Moulines. On upper-confidence-bound policies for switching bandit problems. In *Algorithmic Learning Theory*, 2011.

Hoda Heidari, Michael Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics*, 2012.

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. 2019.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Neural Information Processing Systems*, 2017.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

A Proof of core FEWA guarantees

Lemma 1. *On favorable event ξ_t, if an arm i passes through a filter of window h at round t, the average of its h last pulls cannot deviate significantly from the best available arm $i_t^\star$ at that round, i.e.,*

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

Proof. Let i 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 $\widehat{\mu}_{\max,t}^h$,

$$\overline{\mu}_i^h(N_{i,t}) \geq \widehat{\mu}_i^h(N_{i,t}) - c(h,\delta_t) \geq \widehat{\mu}_{\max,t}^h - 3c(h,\delta_t) \geq \max_{i\in\mathcal{K}_h} \overline{\mu}_i^h(N_{i,t}) - 4c(h,\delta_t), \tag{10}$$

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

$$\widehat{\mu}_{\max,t}^h \geq \widehat{\mu}_i^h(N_{i,t}) \geq \overline{\mu}_i^h(N_{i,t}) - c(h,\delta_t).$$

Second, since the means of arms are decaying, we know that

$$\mu_t^+(\pi_{\mathrm{F}}) \triangleq \mu_{i_t^\star}(N_{i_t^\star,t}) \leq \mu_{i_t^\star}(N_{i_t^\star,t}-1) = \overline{\mu}_{i_t^\star}^1(N_{i_t^\star,t}) \leq \max_{i\in\mathcal{K}} \overline{\mu}_i^1(N_{i,t}) = \max_{i\in\mathcal{K}_1} \overline{\mu}_i^1(N_{i,t}). \tag{11}$$

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

$$\forall h' \leq N_{i,t}-1, \max_{i\in\mathcal{K}_{h'+1}} \overline{\mu}_i^{h'+1}(N_{i,t}) \geq \max_{i\in\mathcal{K}_{h'}} \overline{\mu}_i^{h'}(N_{i,t}).$$

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 $\arg\max_{i\in\mathcal{K}_{h'}} \overline{\mu}_i^{h'}(N_{i,t}) \subseteq \mathcal{K}_{h'+1}$. Let $i_{\max}^{h'} \in \arg\max_{i\in\mathcal{K}_{h'}} \overline{\mu}_i^{h'}(N_{i,t})$. Then for such $i_{\max}^{h'}$, we have

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

where the first and the third inequality are due to confidence bounds on estimates, 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,

$$\max_{i\in\mathcal{K}_{h'}} \overline{\mu}_i^{h'}(N_{i,t}) = \overline{\mu}_{i_{\max}^{h'}}^{h'}(N_{i_{\max}^{h'},t}) \leq \overline{\mu}_{i_{\max}^{h'}}^{h'+1}(N_{i_{\max}^{h'},t}) \leq \max_{i\in\mathcal{K}_{h'+1}} \overline{\mu}_i^{h'+1}(N_{i,t}), \tag{12}$$

because $i_{\max}^{h'} \in \mathcal{K}_{h'+1}$. Gathering Equations 10, 11, and 12 leads to the claim of the lemma,

$$\overline{\mu}_i^h(N_{i,t}) \overset{(10)}{\geq} \max_{i\in\mathcal{K}_h} \overline{\mu}_i^h(N_{i,t}) - 4c(h,\delta_t) \overset{(12)}{\geq} \max_{i\in\mathcal{K}_1} \overline{\mu}_i^1(N_{i,t}) - 4c(h,\delta_t) \overset{(11)}{\geq} \mu_t^+(\pi_{\mathrm{F}}) - 4c(h,\delta_t).$$

□

Corollary 1. *Let $i \in \mathrm{OP}$ be an arm overpulled by* FEWA *at round t and $h_{i,t} \triangleq N_{i,t}^{\pi_{\mathrm{F}}} - N_{i,t}^{\pi^\star} \geq 1$ be the difference in the number of pulls w.r.t. the optimal policy $\pi^\star$ at round t. On favorable event ξ_t, we have*

$$\mu_t^+(\pi_{\mathrm{F}}) - \overline{\mu}_i^{h_{i,t}}(N_{i,t}) \leq 4c(h_{i,t},\sigma,\delta_t). \tag{7}$$

Proof. If i was pulled at round t, then by the condition at Line 10 of Algorithm 1, it means that i passes through all the filters from $h=1$ up to $N_{i,t}$. In particular, since $1 \leq h_{i,t} \leq N_{i,t}$, i passed the filter for $h_{i,t}$, and thus we can apply Lemma 1 and conclude

$$\overline{\mu}_i^h(N_{i,t}) \geq \mu_t^+(\pi_{\mathrm{F}}) - 4c(h_{i,t},\delta_t). \tag{13}$$

□

B Proofs of auxiliary results

Lemma 2. *Let $h^\pi_{i,t} \triangleq |N^\pi_{i,T} - N^{\pi^\star}_{i,T}|$. For any policy π, the regret at round T is no bigger than*

$$R_T(\pi) \leq \sum_{i\in \mathrm{OP}} \sum_{h=0}^{h^\pi_{i,T}-1} \left[\xi_{t^\pi_i(N^\star_{i,T}+h)}\right] \left(\mu^+_T(\pi) - \mu_i(N^{\pi^\star}_{i,T}+h)\right) + \sum_{t=0}^{T} \left[\overline{\xi_t}\right] Lt.$$

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

Proof. We consider the regret at round T. From Equation 3, the decomposition of regret in terms of overpulls and underpulls gives

$$R_T(\pi) = \sum_{i\in \mathrm{UP}} \sum_{t'=N^\pi_{i,T}+1}^{N^{\pi^\star}_{i,T}} \mu_i(t') - \sum_{i\in \mathrm{OP}} \sum_{t'=N^{\pi^\star}_{i,T}+1}^{N^\pi_{i,T}} \mu_i(t').$$

In order to separate the analysis for each arm, we upper-bound all the rewards in the first sum by their maximum $\mu^+_T(\pi) \triangleq \max_{i\in\mathcal{K}} \mu_i(N^\pi_{i,T})$. This upper bound is tight for problem-independent bound because one cannot hope that the unexplored reward would decay to reduce its regret in the worst case. We also notice that there are as many terms in the first double sum (number of underpulls) than in the second one (number of overpulls). This number is equal to $\sum_{\mathrm{OP}} h^\pi_{i,T}$. 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,

$$R_T(\pi) \leq \sum_{i\in \mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \left(\mu^+_T(\pi_\mathrm{F}) - \mu_i(N^{\pi^\star}_{i,T}+t')\right). \tag{14}$$

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

$$\begin{aligned}
R_T(\pi) &\leq \sum_{i\in \mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \sum_{t=0}^{T} \left[t^\pi_i\left(N^{\pi^\star}_{i,T}+t'\right) = t\right] \left(\mu^+_T(\pi) - \mu_i(N^{\pi^\star}_{i,T}+t')\right) \\
&\leq \underbrace{\sum_{i\in \mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \sum_{t=0}^{T} \left[t^\pi_i\left(N^{\pi^\star}_{i,T}+t'\right) = t \wedge \xi_t\right] \left(\mu^+_T(\pi) - \mu_i(N^{\pi^\star}_{i,T}+t')\right)}_{A_\pi} \\
&+ \underbrace{\sum_{i\in \mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \sum_{t=0}^{T} \left[t^\pi_i\left(N^{\pi^\star}_{i,T}+t'\right) = t \wedge \overline{\xi_t}\right] \left(\mu^+_T(\pi) - \mu_i(N^{\pi^\star}_{i,T}+t')\right)}_{B}.
\end{aligned}$$

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

$$A_\pi \leq \sum_{i\in \mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \left[\xi_{t(N^\star_{i,t}+t')}\right] \left(\mu^+_T(\pi) - \mu_i(N^{\pi^\star}_{i,T}+t')\right).$$

For `FEWA`, 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 $\overline{\xi_t}$ 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., $\sum_{i\in\mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \left[t^\pi_i\left(N^{\pi^\star}_{i,T}+t'\right) = t\right] \leq 1$, we get

$$B \leq \sum_{t=0}^{T} \left[\overline{\xi_t}\right] Lt \sum_{i\in \mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \left[t^\pi_i\left(N^{\pi^\star}_{i,T}+t'\right) = t\right] \leq \sum_{t=0}^{T} \left[\overline{\xi_t}\right] Lt.$$

Therefore, we conclude that

$$R_T(\pi) \le \underbrace{\sum_{i\in \text{OP}} \sum_{t'=0}^{h^{\pi}_{i,T}-1} \left[\xi_{t_i^{\pi}(N^{\star}_{i,t}+t')}\right]\left(\mu_T^+(\pi) - \mu_i(N^{\pi^\star}_{i,T}+t')\right)}_{A_\pi} + \underbrace{\sum_{t=0}^{T}\left[\overline{\xi_t}\right] Lt}_{B}.$$

□

Lemma 3. *Let $h_{i,t} \triangleq h^{\pi_{\text{F}}}_{i,t} = |N^{\pi_{\text{F}}}_{i,T} - N^{\pi^\star}_{i,T}|$. For policy π_{F} with parameters (α, δ_0), $A_{\pi_{\text{F}}}$ defined in Lemma 2 is upper-bounded by*

$$\begin{aligned} A_{\pi_{\text{F}}} &\triangleq \sum_{i\in \text{OP}} \sum_{t'=0}^{h_{i,T}-1} \left[\xi_{t_i^{\pi_{\text{F}}}(N^{\star}_{i,t}+t')}\right]\left(\mu_T^+(\pi_{\text{F}}) - \mu_i(N^{\pi^\star}_{i,T}+t')\right) \\ &\le \sum_{i\in \text{OP}_\xi} \left(4\sqrt{2\alpha\sigma^2 \log_+(KT\delta_0^{-1/\alpha})} + 4\sqrt{2\alpha\sigma^2\left(h^{\xi}_{i,T}-1\right)\log_+(KT\delta_0^{-1/\alpha})} + L\right). \end{aligned}$$

Proof. First, we define $h^{\xi}_{i,T} \triangleq \max\left\{h \le h_{i,T} |\ \xi_{t_i^{\pi_{\text{F}}}(N^{\star}_{i,t}+h)}\right\}$, the last overpull of arm i pulled at round $t_i \triangleq t_i^{\pi_{\text{F}}}(N^{\star}_{i,t} + h^{\xi}_{i,T}) \le T$ under ξ_t. Now, we upper-bound $A_{\pi_{\text{F}}}$ by including all the overpulls of arm i until the $h^{\xi}_{i,T}$-th overpull, even the ones under $\overline{\xi_t}$,

$$A_{\pi_{\text{F}}} \triangleq \sum_{i\in \text{OP}} \sum_{t'=0}^{h^{\pi_{\text{F}}}_{i,T}-1} \left[\xi_{t_i^{\pi_{\text{F}}}(N^{\star}_{i,t}+t')}\right]\left(\mu_T^+(\pi_{\text{F}}) - \mu_i(N^{\pi^\star}_{i,T}+t')\right) \le \sum_{i\in \text{OP}_\xi} \sum_{t'=0}^{h^{\xi}_{i,T}-1} \left(\mu_T^+(\pi_{\text{F}}) - \mu_i(N^{\pi^\star}_{i,T}+t')\right),$$

where $\text{OP}_\xi \triangleq \left\{i \in \text{OP} |\ h^{\xi}_{i,T} \ge 1\right\}$. 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 ξ_{t_i}, we apply Corollary 1 to bound the regret caused by previous overpulls of i (possibly none),

$$\begin{aligned} A_{\pi_{\text{F}}} &\le \sum_{i\in \text{OP}_\xi} \mu_T^+(\pi_{\text{F}}) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,T} - 1\right) + 4\left(h^{\xi}_{i,T}-1\right)c\left(h^{\xi}_{i,T}-1, \delta_{t_i}\right) && (15) \\ &\le \sum_{i\in \text{OP}_\xi} \mu_T^+(\pi_{\text{F}}) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,T} - 1\right) + 4\left(h^{\xi}_{i,T}-1\right)c\left(h^{\xi}_{i,T}-1, \delta_{T}\right) && (16) \\ &\le \sum_{i\in \text{OP}_\xi} \mu_T^+(\pi_{\text{F}}) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,T} - 1\right) + 4\sqrt{2\alpha\sigma^2\left(h^{\xi}_{i,T}-1\right)\log_+\left(KT\delta_0^{-1/\alpha}\right)}, && (17) \end{aligned}$$

with $\log_+(x) \triangleq \max(\log(x), 0)$. 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 4 with $\log_+(KT^\alpha) \le \alpha \log_+(KT)$ for $\alpha > 1$. If $N^{\pi^\star}_{i,T} = 0$ and $h^{\xi}_{i,T} = 1$ then

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

since and $\mu^+(\pi_{\text{F}}) \le L$ and $\mu_i(0) \ge 0$ by the assumptions of our setting. Otherwise, we can decompose

$$\mu_T^+(\pi_{\text{F}}) - \mu_i(N^{\pi^\star}_{i,T} + h^{\xi}_{i,T} - 1) = \underbrace{\mu_T^+(\pi_{\text{F}}) - \mu_i(N^{\pi^\star}_{i,T} + h^{\xi}_{i,T} - 2)}_{A_1} + \underbrace{\mu_i(N^{\pi^\star}_{i,T} + h^{\xi}_{i,T} - 2) - \mu_i(N^{\pi^\star}_{i,T} + h^{\xi}_{i,T} - 1)}_{A_2}.$$

For term A_1, since arm i was overpulled at least once by `FEWA`, it passed at least the first filter. Since this $h^{\xi}_{i,T}$-th overpull is done under ξ_{t_i}, by Lemma 1 we have that

$$A_1 \le 4c(1, \delta_{t_i}) \le 4c(1, K^{-1}T^{-\alpha}) \le 4\sqrt{2\alpha\sigma^2 \log_+\left(KT\delta_0^{-1/\alpha}\right)}.$$

The second difference, $A_2 = \mu_i(N_{i,T}^{\pi^\star} + h_{i,T}^{\xi} - 2) - \mu_i(N_{i,T}^{\pi^\star} + h_{i,T}^{\xi} - 1)$ cannot exceed L, since by the assumptions of our setting, the maximum decay in one round is bounded. Therefore, we further upper-bound Equation 17 as

$$A_{\pi_F} \leq \sum_{i \in \mathrm{OP}_\xi} \left(4\sqrt{2\alpha\sigma^2 \log_+\left(KT\delta_0^{-1/\alpha}\right)} + 4\sqrt{2\alpha\sigma^2\left(h_{i,T}^{\xi} - 1\right)\log_+\left(KT\delta_0^{-1/\alpha}\right)} + L \right). \tag{18}$$

□

Lemma 4. *Let $\zeta(x) = \sum_n n^{-x}$. Thus, with $\delta_t = \delta_0/(Kt^\alpha)$ and $\alpha > 4$, we can use Proposition 4 and get*

$$\mathbb{E}[B] \triangleq \sum_{t=0}^{T} p(\overline{\xi_t}) Lt \leq \sum_{t=0}^{T} \frac{Lt\delta_0}{2t^{\alpha-2}} \leq L\delta_0 \frac{\zeta(\alpha-3)}{2}.$$

C Minimax regret analysis of FEWA

Theorem 1. *For any rotting bandit scenario with means $\{\mu_i(n)\}_{i,n}$ satisfying Assumption 1 with bounded decay L and any time horizon T, FEWA run with $\alpha = 5$, $\delta_0 = 1$, i.e., with $\delta_t = 1/(Kt^5)$, suffers an expected regret*[9]

$$\mathbb{E}[R_T(\pi_F)] \leq 13\sigma(\sqrt{KT} + K)\sqrt{\log(KT)} + KL.$$

Proof. To get the problem-independent upper bound for FEWA, we need to upper-bound the regret by quantities which do not depend on $\{\mu_i\}_i$. The proof is based on Lemma 2, where we bound the expected values of terms A_{π_F} and B from the statement of the lemma. We start by noting that on high-probability event ξ_T, we have by Lemma 3 and $\alpha = 5$ that

$$A_{\pi_F} \leq \sum_{i \in \mathrm{OP}_\xi} \left(4\sqrt{10\sigma^2 \log(KT)} + 4\sqrt{10\sigma^2(h_i - 1)\log(KT)} + L \right).$$

Since $\mathrm{OP}_\xi \subseteq \mathrm{OP}$ and there are at most $K-1$ overpulled arms, we can upper-bound the number of terms in the above sum by $K-1$. Next, the total number of overpulls $\sum_{i \in \mathrm{OP}} h_{i,T}$ 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., $h_{i,T} = T/(K-1)$,

$$\begin{aligned} A_{\pi_F} &\leq (K-1)(4\sqrt{10\sigma^2 \log(KT)} + L) + 4\sqrt{10\sigma^2 \log(KT)} \sum_{i \in \mathrm{OP}} \sqrt{(h_{i,T} - 1)} \\ &\leq (K-1)(4\sqrt{10\sigma^2 \log(KT)} + L) + 4\sqrt{10\sigma^2(K-1)T \log(KT)}. \end{aligned} \tag{19}$$

Now, we consider the expectation of term B from Lemma 2. According to Lemma 4, with $\alpha = 5$ and $\delta_0 = 1$,

$$\mathbb{E}[B] \leq \frac{L\zeta(2)}{2} = \frac{L\pi^2}{12}. \tag{20}$$

Therefore, using Lemma 2 together with Equations 19 and 20, we bound the total expected regret as

$$\mathbb{E}[R_T(\pi_F)] \leq 4\sqrt{10\sigma^2(K-1)T \log(KT)} + (K-1)(4\sqrt{10\sigma^2 \log(KT)} + L) + \frac{L\pi^2}{6}. \tag{21}$$

□

Corollary 3. FEWA run with $\alpha > 3$ and $\delta_0 \triangleq 2\delta/\zeta(\alpha-2)$ achieves with probability $1-\delta$,

$$R_T(\pi_F) = A_{\pi_F} \leq 4\sqrt{2\alpha\sigma^2 \log_+\left(\frac{KT}{\delta_0^{1/\alpha}}\right)}\left(K - 1 + \sqrt{(K-1)T}\right) + (K-1)L.$$

[9] See Corollary 3 and 4 for the high-probability result.

Proof. We consider the event $\bigcup_{t\leq T}\xi_t$ which happens with probability

$$1-\sum_{t\leq T}\frac{Kt^2\delta_t}{2}\leq 1-\sum_{t\leq T}\frac{Kt^2\delta_t}{2}\leq 1-\frac{\zeta(\alpha-2)\delta_0}{2}.$$

Therefore, by setting $\delta_0\triangleq 2\delta/\zeta(\alpha-2)$, we have that $B=0$ with probability $1-\delta$ since $\left[\overline{\xi_t}\right]=0$ for all t. We can then use the same analysis of $A_{\pi_{\mathrm{F}}}$ as in Theorem 1 to get

$$R_T(\pi_{\mathrm{F}})=A_{\pi_{\mathrm{F}}}\leq 4\sqrt{2\alpha\sigma^2\log_+\left(\frac{KT}{\delta_0^{1/\alpha}}\right)}\left(K-1+\sqrt{(K-1)T}\right)+(K-1)L.$$

□

D Problem-dependent regret analysis of FEWA

Lemma 5. *$A_{\pi_{\mathrm{F}}}$ defined in Lemma 2 is upper-bounded by a problem-dependent quantity,*

$$A_{\pi_{\mathrm{F}}}\leq\sum_{i\in\mathcal{K}}\left(\frac{32\alpha\sigma^2\log_+(KT\delta_0^{-1/\alpha})}{\Delta_{i,h^+_{i,T}-1}}+\sqrt{32\alpha\sigma^2\log_+(KT\delta_0^{-1/\alpha})}\right)+(K-1)L.$$

Proof. We start from the result of Lemma 3,

$$A_{\pi_{\mathrm{F}}}\leq\sum_{i\in\mathrm{OP}_\xi}\left(4\sqrt{2\alpha\sigma^2\log(KT\delta_0^{-1/\alpha})}\left(1+\sqrt{h^\xi_{i,T}-1}\right)\right)+(K-1)L. \tag{22}$$

We want to bound $h^\xi_{i,T}$ with a problem dependent quantity $h^+_{i,T}$. We remind the reader that for arm i at round T, the $h^\xi_{i,T}$-th overpull has been on ξ_{t_i} pulled at round t_i. Therefore, Corollary 1 applies and we have

$$\begin{aligned}\overline{\mu}_i^{h^\xi_{i,T}-1}\left(N^{\pi^\star}_{i,T}+h^\xi_{i,T}-1\right)&\geq\mu^+_T(\pi_{\mathrm{F}})-4c\left(h^\xi_{i,T}-1,\delta_{t_i}\right)\geq\mu^+_T(\pi_{\mathrm{F}})-4c\left(h^\xi_{i,T}-1,\delta_T\right)\\&\geq\mu^+_T(\pi_{\mathrm{F}})-4\sqrt{\frac{2\alpha\sigma^2\log\left(KT\delta_0^{-1/\alpha}\right)}{h^\xi_{i,T}-1}}\geq\mu^-_T(\pi^\star)-4\sqrt{\frac{2\alpha\sigma^2\log\left(KT\delta_0^{-1/\alpha}\right)}{h^\xi_{i,T}-1}},\end{aligned}$$

with $\mu^-_T(\pi^\star)\triangleq\min_{i\in\mathcal{K}}\mu_i\left(N^\star_{i,T}-1\right)$ being the lowest mean reward for which a noisy value was ever obtained by the optimal policy. $\mu^-_T(\pi^\star)<\mu^+_T(\pi_{\mathrm{F}})$ implies that the regret is 0. Indeed, in that case the next possible pull with the largest mean for π_{F} is *strictly larger* than the mean of the last pull for $\pi^\star$. Thus, there is no underpull at this round for π_{F} and $R_T(\pi_{\mathrm{F}})=0$ according to Equation 3. Therefore, we can assume $\mu^-_T(\pi^\star)\geq\mu^+_T(\pi_{\mathrm{F}})$ for the regret bound. Next, we define $\Delta_{i,h}\triangleq\mu^-_T(\pi^\star)-\overline{\mu}_i^h\left(N^\star_{i,t}+h\right)$ as 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. Thus, we have the following bound for $h^\xi_{i,T}$,

$$h^\xi_{i,T}\leq 1+\frac{32\alpha\sigma^2\log\left(KT\delta_0^{-1/\alpha}\right)}{\Delta_{i,h^\xi_{i,T}-1}}.$$

Next, $h^\xi_{i,T}$ has to be smaller than the maximum such h, for which the inequality just above is satisfied if we replace $h^\xi_{i,T}$ by h. Therefore,

$$h^\xi_{i,T}\leq h^+_{i,T}\triangleq\max\left\{h\leq T\middle|\ h\leq 1+\frac{32\alpha\sigma^2\log_+\left(KT\delta_0^{-1/\alpha}\right)}{\Delta^2_{i,h-1}}\right\}. \tag{23}$$

Since the square-root function is increasing, we can upper-bound Equation 17 by replacing $h^{\xi}_{i,T}$ by its upper bound $h^{+}_{i,T}$ to get

$$A_{\pi_{\mathrm{F}}} \le \sum_{i\in \mathrm{OP}_\xi} \left(4\sqrt{2\alpha\sigma^2 \log_+(KT\delta_0^{-1/\alpha})}\left(1+\sqrt{h^+_{i,T}-1}\right)+L\right)$$
$$\le \sum_{i\in \mathrm{OP}_\xi} \left(\sqrt{32\alpha\sigma^2 \log_+(KT\delta_0^{-1/\alpha})}\left(1+\frac{\sqrt{32\alpha\sigma^2 \log_+(KT\delta_0^{-1/\alpha})}}{\Delta_{i,h^+_{i,T}-1}}\right)+L\right).$$

The quantity OP_ξ is depends on the execution. Notice that there are at most $K-1$ arms in OP_ξ and that $\mathrm{OP}\subset\mathcal{K}$. Therefore, we have

$$A_{\pi_{\mathrm{F}}} \le \sum_{i\in\mathcal{K}} \left(\frac{32\alpha\sigma^2 \log_+\left(KT\delta_0^{-1/\alpha}\right)}{\Delta_{i,h^+_{i,T}-1}} + \sqrt{32\alpha\sigma^2 \log_+\left(KT\delta_0^{-1/\alpha}\right)}\right) + (K-1)L.$$

□

Corollary 2 (problem-dependent guarantee)**.** *For $\delta_t \triangleq 1/(Kt^5)$, the regret is bounded as*

$$\mathbb{E}[R_T(\pi_{\mathrm{F}})] \le \sum_{i\in\mathcal{K}}\left(\frac{C_5\log(KT)}{\Delta_{i,h^+_{i,T}-1}} + \sqrt{C_5\log(KT)} + L\right) \text{ with } C_\alpha \triangleq 32\alpha\sigma^2 \text{ and } h^+_{i,T} \text{ defined in Equation 9.}$$

Proof. Using Lemmas 2, 4, and 5 we get

$$\mathbb{E}[R_T(\pi_{\mathrm{F}})] = \mathbb{E}[A_{\pi_{\mathrm{F}}}] + \mathbb{E}[B] \le \sum_{i\in\mathcal{K}}\left(\frac{32\alpha\sigma^2\log(KT)}{\Delta_{i,h^+_{i,T}-1}} + \sqrt{32\alpha\sigma^2\log(KT)}\right) + (K-1)L + \frac{L\pi^2}{6}$$
$$\le \sum_{i\in\mathcal{K}}\left(\frac{32\alpha\sigma^2\log(KT)}{\Delta_{i,h^+_{i,T}-1}} + \sqrt{32\alpha\sigma^2\log(KT)} + L\right).$$

□

Corollary 4. `FEWA` run with $\alpha > 3$ and $\delta_0 \triangleq 2\delta/\zeta(\alpha-2)$ achieves with probability $1-\delta$,

$$R_T(\pi_{\mathrm{F}}) \le \sum_{i\in\mathcal{K}}\left(\frac{32\alpha\sigma^2\log_+\left(\frac{KT\zeta(\alpha-2)^{1/\alpha}}{(2\delta)^{1/\alpha}}\right)}{\Delta_{i,h^+_{i,T}-1}} + \sqrt{32\alpha\sigma^2\log_+\left(\frac{KT\zeta(\alpha-2)^{1/\alpha}}{(2\delta)^{1/\alpha}}\right)}\right) + (K-1)L.$$

Proof. We consider the event $\cup_{t\le T}\xi_t$ which happens with probability

$$1 - \sum_{t\le T}\frac{Kt^2\delta_t}{2} \le 1 - \sum_{t\le T}\frac{Kt^2\delta_t}{2} \le 1 - \frac{\zeta(\alpha-2)\delta_0}{2}.$$

Therefore, by setting $\delta_0 \triangleq 2\delta/\zeta(\alpha-2)$, we have that with probability $1-\delta$, $B=0$ since $\left[\overline{\xi_t}\right]=0$ for all t. We use Lemma 5 to get the claim of the corollary. □

E Efficient algorithm `EFF-FEWA`

In Algorithm 3, we present `EFF-FEWA`, an algorithm that stores at most $2K\log_2(t)$ of statistics. More precisely, for $j \le \log_2(N^{\pi_{\mathrm{EF}}}_{i,t})$, we let $\widehat{s}^p_{i,j}$ and $\widehat{s}^c_{i,j}$ be the current and pending j-th statistic for arm i. We then present an analysis of `EFF-FEWA`.

Algorithm 3 `EFF-FEWA`

Input: $\mathcal{K}$, δ_0, α

pull each arm once, collect reward, and initialize $N_{i,K} \leftarrow 1$
for $t \leftarrow K+1, K+2, \ldots$ **do**
  $\delta_t \leftarrow \delta_0/(Kt^\alpha)$
  $j \leftarrow 0$ {*initialize bandwidth*}
  $\mathcal{K}_1 \leftarrow \mathcal{K}$ {*initialize with all the arms*}
  $i(t) \leftarrow$ `none`
  while $i(t)$ is `none` **do**
    $\mathcal{K}_{2^{j+1}} \leftarrow$ `EFF_Filter`$(\mathcal{K}_{2^j}, j, \delta_t)$
    $j \leftarrow j+1$
    if $\exists i \in \mathcal{K}_{2^j}$ such that $N_{i,t} \leq 2^j$ **then**
      $i(t) \leftarrow i$
    end if
  end while
  receive $r_i(N_{i,t+1}) \leftarrow r_{i(t),t}$
  `EFF_Update`$(i(t), r_i(N_{i,t+1}), t+1)$
end for

Algorithm 4 `EFF_Filter`

Input: $\mathcal{K}_{2^j}$, j, δ_t, σ

$c(2^j, \delta_t) \leftarrow \sqrt{2\sigma^2/2^j \log \delta_t^{-1}}$
$\widehat{s}^c_{\max,j} \leftarrow \max_{i \in \mathcal{K}_h} \widehat{s}^c_{i,j}$
for $i \in \mathcal{K}_h$ **do**
  $\Delta_i \leftarrow \widehat{s}^c_{\max,j} - \widehat{s}^c_{i,j}$
  if $\Delta_i \leq 2c(2^j, \delta_t)$ **then**
    add i to $\mathcal{K}_{2^{j+1}}$
  end if
end for

Output: $\mathcal{K}_{2^{j+1}}$

Algorithm 5 `EFF_Update`

Input: i, r, t

$N_{i(t),t} \leftarrow N_{i(t),t-1} + 1$
$R_i^{\text{total}} \leftarrow R_i^{\text{total}} + r$ {*keep track of total reward*}
if $\exists j$ such that $N_{i,t} = 2^j$ **then**
  $\widehat{s}^c_{i,j} \leftarrow R_i^{\text{total}}/N_{i,t}$ {*initialize new statistics*}
  $\widehat{s}^p_{i,j} \leftarrow 0$
  $n_{i,j} \leftarrow 0$
end if
for $j \leftarrow 0 \ldots \log_2(N_{i,t})$ **do**
  $n_{i,j} \leftarrow n_i + 1$
  $\widehat{s}^p_{\max,j} \leftarrow \widehat{s}^p_{\max,j} + r$
  if $n_{i,j} = 2^j$ **then**
    $\widehat{s}^c_{\max,j} \leftarrow \widehat{s}^p_{\max,j}/2^j$
    $n_{i,j} \leftarrow 0$
    $\widehat{s}^p_{\max,j} \leftarrow 0$
  end if
end for

On one hand, at any time t, $\widehat{s}^c_{i,j}$ is the average of 2^{j-1} consecutive reward samples for arm i within the last $2^j - 1$ sample. These statistics are used in the filtering process as they are representative of exactly 2^{j-1} recent samples. On the other hand, $\widehat{s}^p_{i,j}$ stores the pending samples that are not yet taken into account by $\widehat{s}^c_{i,j}$. Therefore, each time we pull arm i, we update all the pending averages. When

the pending statistic is the average of the 2^{j-1} last samples then we set $\widehat{s}^c_{i,j} \leftarrow \widehat{s}^p_{i,j}$ and we reinitialize $\widehat{s}^p_{i,j} \leftarrow 0$.

How does that modify Lemma 1? We let $\overline{\mu}_i^{h',h''}$ be the average of the samples between the h'-th last one and the h''-th last one (included) with $h'' > h'$. `FEWA` was controlling $\overline{\mu}_i^{1,h}$ for each arm, `EFF-FEWA` controls $\overline{\mu}_i^{h'_i,h'_i+2^{j-1}}$ with different $h'_i \leq 2^{j-1}-1$ for each arm. However, since the means of arms are non-increasing, we can consider the worst case when the arm with the highest mean available at that round is estimated on its last samples (the smaller one) and the bad arms are estimated on their oldest possibles samples (the larger one).

Lemma 6. *On the favorable event ξ_t, if an arm i passes through a filter of window h at round t, the average of its h last pulls cannot deviate significantly from the best available arm $i^\star_t$ at that round,*

$$\overline{\mu}_i^{2^{j-1},2^j-1} \geq \mu_t^+(\pi_{\mathrm{F}}) - 4c(h,\delta_t).$$

Then, we modify Corollary 1 to have the following efficient version of it.

Corollary 5. *Let $i \in \mathrm{OP}$ be an arm overpulled by `EFF-FEWA` at round t and $h^{\pi_{\mathrm{EF}}}_{i,t} \triangleq N^{\pi_{\mathrm{EF}}}_{i,t} - N^{\pi^\star}_{i,t} \geq 1$ be the difference in the number of pulls w.r.t. the optimal policy $\pi^\star$ at round t. On the favorable event ξ_t, we have that*

$$\mu_t^+(\pi_{\mathrm{EF}}) - \overline{\mu}^{h^{\pi_{\mathrm{EF}}}_{i,t}}(N_{i,t}) \leq \frac{4\sqrt{2}}{\sqrt{2}-1} c(h^{\pi_{\mathrm{EF}}}_{i,t}, \delta_t).$$

Proof. If i was pulled at round t, then by the condition at Line 10 of Algorithm 3, it means that i passes through all the filters until at least window 2^f such that $2^f \leq h^{\pi_{\mathrm{EF}}}_{i,t} < 2^{f+1}$. Note that for $h^{\pi_{\mathrm{EF}}}_{i,t} = 1$, then `EFF-FEWA` has the same guarantee as `FEWA` since the first filter is always up to date. Then for $h^{\pi_{\mathrm{EF}}}_{i,t} \geq 2$,

$$\begin{aligned}
\overline{\mu}_i^{1,h^{\pi_{\mathrm{EF}}}_{i,t}}(N_{i,t}) \geq \overline{\mu}_i^{1,2^f-1}(N_{i,t}) &= \frac{\sum_{j=1}^f 2^{j-1}\overline{\mu}_i^{2^{j-1},2^j-1}}{2^f-1} && (24)\\
&\geq \mu_t^+(\pi_{\mathrm{EF}}) - \frac{4\sum_{j=1}^f 2^{j-1}c(2^{j-1},\delta)}{2^f-1} = \mu_t^+(\pi_{\mathrm{EF}}) - 4c(1,\delta_t)\frac{\sum_{j=1}^f \sqrt{2}^{j-1}}{2^f-1} && (25)\\
&= \mu_t^+(\pi_{\mathrm{EF}}) - 4c(1,\delta_t)\frac{\sqrt{2}^f-1}{(2^f-1)(\sqrt{2}-1)} \geq \mu_t^+(\pi_{\mathrm{EF}}) - 4c(1,\delta_t)\frac{1}{\sqrt{2}^f(\sqrt{2}-1)} && (26)\\
&= \mu_t^+(\pi_{\mathrm{EF}}) - \frac{4\sqrt{2}}{\sqrt{2}-1}c\big(2^{f+1},\delta_t\big) \geq \mu_t^+(\pi_{\mathrm{EF}}) - \frac{4\sqrt{2}}{\sqrt{2}-1}c\big(h^{\pi_{\mathrm{EF}}}_{i,t},\delta_t\big), && (27)
\end{aligned}$$

where Equation 24 uses that the average of older means is larger than average of the more recent ones and then decomposes 2^f-1 means onto a geometric grid. Then, Equation 25 uses Lemma 6 and make the dependence of $c(2^{j-1},\delta)$ on j explicit. Next, Equations 26 and 27 use standard algebra to derive a lower bound and that $c(h,\delta)$ decreases with h. □

Armed with the above, we use the same proof as the one we have for `FEWA` and derive minimax and problem-dependent upper bounds for `EFF-FEWA` using Corollary 5 instead of Corollary 1.

Corollary 6 (minimax guarantee for `EFF-FEWA`)**.** *For any rotting bandit scenario with means $\{\mu_i(n)\}_{i,n}$ satisfying Assumption 1 with bounded decay L and any time horizon T, `EFF-FEWA` with $\delta_t = 1/(Kt^5)$, $\alpha = 5$, and $\delta_0 = 1$, has its expected regret upper-bounded as*

$$\mathbb{E}[R_T(\pi_{\mathrm{EF}})] \leq 13\sigma\left(\frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{KT}+K\right)\sqrt{\log(KT)} + KL.$$

Corollary 7 (problem-dependent guarantee for `EFF-FEWA`)**.** *For $\delta_t = 1/(Kt^5)$, the regret of `EFF-FEWA` is upper-bounded as*

$$R_T(\pi_{\mathrm{EF}}) \leq \sum_{i\in\mathcal{K}}\left(\frac{C_5\frac{2}{3-2\sqrt{2}}\log(KT)}{\Delta_{i,h^+_{i,T}-1}} + \sqrt{C_5\log(KT)} + L\right),$$

with $C_\alpha \triangleq 32\alpha\sigma^2$ and $h^+_{i,T}$ defined in Equation 9.