Rotting Bandits Are No Harder Than Stochastic Ones

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, Michal Valko
Lelivrescolaire.fr Otto-von-Guericke-Universität FAIR INRIA Lille

Abstract

In stochastic multi-armed bandit (MAB), the reward distribution of each arm is assumed to be stationary. This assumption is often violated in practice (e.g., in recommendation systems), where the reward of an arm may change whenever is selected (i.e., rested bandit setting). In this paper, we consider the *non-parametric rotting bandit* setting, where rewards can only decrease. We introduce the *filtering on expanding window average* (FEWA) algorithm that constructs moving averages of increasing windows to identify arms that are more likely to return high rewards when pulled once more. We prove that for an unknown horizon T, and without any knowledge on the decreasing behavior of the K arms, FEWA achieves problem-dependent, $\widetilde{\mathcal{O}}(\log(KT))$, and problem-independent, $\widetilde{\mathcal{O}}(\sqrt{KT})$, regret bounds. This result substantially improves over the algorithm proposed by Levine et al. (2017), which suffers regret $\widetilde{\mathcal{O}}(K^{1/3}T^{2/3})$, and it matches standard bounds for the stochastic MAB setting, thus showing that the rotting bandit is not harder. Finally, we report simulations confirming the theoretical improvements of FEWA.

1 Introduction

The multi-arm bandit (MAB) framework (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2019) formalizes 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 reward. In the *stochastic setting* (Thompson, 1933; Auer et al., 2002a), each arm is characterized by a stationary reward distribution and whenever an arm is pulled, an i.i.d. sample from the corresponding distribution is observed. Despite the extensive algorithmic and theoretical study of this setting, the stationarity assumption is often too restrictive in practice (e.g., the preferences of users may change over time). 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 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. These settings fall into the so-called *restless* bandit scenario, where the arms' value evolves *independently* from the decisions of the agent. On the other hand, in many problems, the value of an arm changes only when it is pulled (i.e., the *rested* bandit scenario). For instance, the value of a service may deteriorate only when it is actually used (e.g., if a recommender system shows always the same item to the users, they may get bored (Warlop et al., 2018)). Similarly, 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 may decrease whenever pulled. Heidari et al. (2016) studied this problem when rewards 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

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

known up to a constant factor. Finally, Levine et al. (2017) defined both non-parametric and parametric noisy rotting bandits, for which they derive algorithms with regret guarantees. 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 constructs moving average estimates of increasing 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, we show that FEWA achieves a regret of $\widetilde{\mathcal{O}}(\sqrt{KT})$, 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 stationary case. When rewards are constant we also recover *standard problem-dependent regret guarantees* (up to constants), while in the rotting bandit scenario with no noise, the regret reduces to the one of Heidari et al. (2016). Numerical simulations confirm our theoretical results and show the superiority of FEWA over wSWA.

2 Preliminaries

We consider a rotting bandit scenario similar to (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 $r_{i(t),t}$. The reward associated to each arm i is a σ^2-sub-Gaussian r.v. with expected value $\mu_i(n)$, which depends on the number of times n it was pulled before ($\mu_i(0)$ is the initial expected value).[1] 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 rewards observed until round t, 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. We use $r_i(n)$ to denote the random reward of arm i when pulled for the $n+1$-th time, i.e., $r_{i(t),t} = r_{i(t)}(N_{i(t),t})$. We introduce a non-parametric rotting assumption with bounded decay.

Assumption 1. *The reward functions μ_i are non-increasing with bounded decays $-L \leq \mu_i(n+1) - \mu_i(n) \leq 0$. The initial expected value is bounded as $\mu_i(0) \in [0, L]$. We refer to this set of functions as $\mathcal{L}_L$.*

[1] Our definition slightly differs from Levine et al. (2017). Here $\mu_i(n)$ denotes the expected value of arm i *after n pulls* instead of when it is pulled *for the n-th time*.

The learning problem. A learning policy π is a function from the history of observations to arms, i.e., $\pi(\mathcal{H}_t) \in \mathcal{K}$. In the following, we often use $\pi(t) \stackrel{\text{def}}{=} \pi(\mathcal{H}_t)$. 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 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 (oracle) policy.

Proposition 1 (Heidari et al. (2016))**.** *If the expected value of each arm $\{\mu_i(n)\}_{i,n}$ is known, the 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 with performance as close to $\pi^\star$'s as possible. We define the (random) regret as

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

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 and definition of regret reduce to standard stochastic bandits.

Let $N^\star_{i,T}$ be the (deterministic) number of times that arm i is pulled by the oracle policy $\pi^\star$ up to time T (excluded). Similarly, for a policy π, let $N^\pi_{i,T}$ be the (random) number pulls of arm i. 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

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

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

Regret bounds. We report existing regret bounds for two special cases. We start with the minimax regret lower bound for stochastic bandits.

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 σ-sub-Gaussian arms such that π suffers a regret*

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

where the expectation is w.r.t. both the randomization over rewards and algorithm's internal randomization.

Heidari et al. (2016) derived regret lower and upper bounds for 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 the greedy policy that selects at each round the arm with the largest reward observed so far, i.e., $\pi^{\sigma_0}(t) = \arg\max_i(\mu_i(N_{i,t}-1))$. For any deterministic rotting bandits (i.e., $\sigma = 0$) satisfying Assumption 1 with bounded decay L, π^{σ_0} suffers an expected regret

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

Any problem in these two settings above is a rotting problem with parameters (σ, L). Therefore, the performance of any algorithm on the general rotting problem is also bounded by two 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 is that we cannot rely on all samples observed until time t to predict which arm is likely to return the highest reward in the future. In fact, the older a sample, the less representative for future rewards. This suggests constructing estimates using the more recent samples. Nonetheless, discarding older rewards reduces the number of samples used in the estimates, thus increasing their variance. In Alg. 1 we introduce 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.

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: $\quad\quad \mathcal{K}_{h+1} \leftarrow \text{FILTER}(\mathcal{K}_h, h, \delta_t)$
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: $\quad\quad\quad i(t) \leftarrow \arg\min_{i \in \mathcal{K}_h} N_{i,t}$
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

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

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

We first describe the subroutine FILTER in Alg. 2, which receives as input a set of active arms $\mathcal{K}_h$, a window h, and a confidence parameter δ, and returns 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 $h \leq n$ most recent rewards observed from i. The subroutine FILTER 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 Prop. 4).

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, it should be directly discarded. 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., as few samples are used in constructing $\widehat{\mu}_i^h(N_{i,t})$, the estimation error may be high). Thus, FEWA keeps refining $\mathcal{K}_h$ for larger windows in the attempt of constructing 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 (recall that $\widehat{\mu}_i^h(N_{i,t})$ is not defined for $h > N_{i,t}$). Finally, FEWA 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.

Runtime and memory usage. At each round t, FEWA 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 incrementally at a cost $\mathcal{O}(1)$, the worst-case time and memory complexity per round is $\mathcal{O}(Kt)$, which amount to a total $\mathcal{O}(KT^2)$ cost, which may not be practical for large T.[2]

In App. E we detail EFF-FEWA, an efficient variant of FEWA. EFF-FEWA is built around two main ideas.[3] First, at any time t we can avoid calling FILTER for all possible windows h starting from 1 with an increment of 1. In fact, the confidence interval $c(h, \sigma, \delta_t)$ decreases as $1/\sqrt{h}$ and we could select windows h with an exponential increment so that confidence intervals between two consecutive calls to FILTER have a constant ratio. In practice, we replace the window increment (Line 9 of FEWA) by a geometric window $h = 2^j$. This modification alone is not enough to reduce the amount of computation. While we reduce the number of estimates that we construct, updating $\widehat{\mu}_i^h$ from $h = 2^j$ to $h = 2^{j+1}$ still requires spanning over past samples, thus leading to the same $\mathcal{O}(Kt)$ complexity in the worst-case. In order to reduce the overall complexity, we avoid recomputing $\widehat{\mu}_i^h$ at each call of FILTER and we replace it with *precomputed* estimates. Whenever $N_{i,t} = 2^j$ for some j, we create an estimate $\widehat{s}_{i,j}^c$ by averaging all the last $N_{i,t}$ samples. These estimates are then used whenever FILTER is called with $h = 2^j$. Instead of updating $\widehat{s}_{i,j}^c$ at each new sample, we create an associated *pending* estimate $\widehat{s}_{i,j}^p$ which averages all the more recent samples. More formally, let t be the time when $N_{i,t} = 2^j$, then $\widehat{s}_{i,j}^p$ is initialized at 0 and it then stores the average of all the samples observed from t to t', when $N_{i,t'} = 2^{j+1}$ (i.e., $\widehat{s}_{i,j}^p$ is averaging at most 2^j samples). At this point, the 2^j samples averaged in $\widehat{s}_{i,j}^c$ are *outdated* and they are replaced by the new average $\widehat{s}_{i,j}^p$, which is then reinitialized to 0. The sporadic update of the precomputed estimates and the small number of them allows to drastically reduce per-round time and space complexity to $\mathcal{O}(K \log t)$. Furthermore, EFF-FEWA preserves the same regret guarantees as FEWA. In the worst case, $\widehat{s}_{i,j}^c$ may not cover the last $2^{j-1} - 1$ samples, which can make it quite inaccurate. Nonetheless, the precomputed estimates with smaller windows (i.e., $j' < j$) are updated more frequently, thus effectively covering the $2^{j-1} - 1$ samples "missed" by $\widehat{s}_{i,j}^c$. As a result, the active sets returned by FILTER are still accurate enough to derive regret guarantees that are only a constant factor worse than FEWA.

4 Regret Analysis

We first state the major theoretical result of the paper, a problem-independent regret bound for FEWA and sketch its proof in Sect. 4.1. Then, we derive problem-dependent guarantees in Sect. 4.2.

Theorem 1. *For any rotting bandit scenario with means $\{\mu_i(n)\}_{i,n}$ satisfying Asm. 1 with bounded decay L and any time horizon T,* FEWA *run with $\alpha = 5$ and $\delta_t = 1/(Kt^5)$, suffers an expected regret*[4]

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

Comparison to Levine et al. (2017). The regret of wSWA is bounded by $\widetilde{\mathcal{O}}(\mu_{\max}^{1/3} K^{1/3} T^{2/3})$ for rotting functions bounded in $[0, \mu_{\max}]$. In our setting, we do not restrict rewards to stay positive but we bound the per-round decay by L, thus leading to rotting functions bounded in $[-LT, L]$. As a result, when applying wSWA to our setting, we should set $\mu_{\max} = L(T+1)$, which leads to $\mathcal{O}(T)$ regret, thus showing that according to its original analysis wSWA may not be able to learn in our general setting. On the other hand, we could use FEWA in the setting of Levine et al. (2017) by setting $L = \mu_{\max}$ as the largest drop that could occur. In this case, FEWA suffers a regret of $\widetilde{\mathcal{O}}(\sqrt{KT})$, thus significantly improving over wSWA in this setting

[2] This analysis is worst-case. In many cases, the number of samples for the suboptimal arms may be much smaller than $\mathcal{O}(t)$. For instance, in stochastic bandits it is as little as $\mathcal{O}(\log t)$, thus reducing the complexity to $\mathcal{O}(KT \log T)$.

[3] A similar yet different approach has appeared independently in the context of streaming mining (Bifet and Gavaldà, 2007).

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

as well. The improvement is mostly due to the fact that FEWA exploits filters using moving averages with increasing windows to discard arms that are suboptimal w.h.p. 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. Moreover, FEWA in anytime, while the fixed exploratory phase of wSWA requires either to know T or to resort to a doubling trick, which often performs poorly in practice.

Comparison to deterministic rotting bandit. 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 terms depending on the noise level σ are separated from the terms depending on the rotting level L, while in wSWA these are coupled in a $L^{1/3}\sigma^{2/3}$ factor in the leading term.

Comparison to stochastic bandit. The regret of FEWA 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 needs confidence bounds to hold for different windows at the same time, which requires an additional union bound and thus larger confidence intervals w.r.t. UCB1. On the other hand, our worst-case analysis shows that some of the difficult problems that reach the worst-case bound of Thm. 1 are realized with constant functions, which is the standard stochastic bandits, for which MOSS-like (Audibert and Bubeck, 2009) algorithms achieve regret guarantees without the $\log T$ factor. Thus, 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

We provide 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 $1 \leq h \leq n$, 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 at round t, if 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 introduce Hoeffding's concentration inequality and the high probability event that we leverage in the analysis.

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

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

For any round t and confidence $\delta_t \triangleq \delta_0/(Kt^\alpha)$, let

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

be the event under which the estimates constructed by FEWA *at round t are all accurate up to $c(h,\delta_t)$. Taking a union bound gives $\mathbb{P}(\xi_t) \geq 1 - Kt^2\delta_t/2$.*

Active set. We derive a crucial lemma that provides support to the arm selection process obtained by a series of refinements through 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

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

We denote by $\mu_t^+(\pi_F) \triangleq \max_{i\in\mathcal{K}} \mu_i(N_{i,t}^{\pi_F})$, the reward obtained by pulling $i_t^\star$. The dependence on π_F in the definition of $\mu_t^+(\pi_F)$ stresses the fact that we consider what the oracle policy would do at the state reached by π_F. While FEWA cannot directly match the performance of the oracle arm, the following lemma shows that reward averaged over the last h pulls of any arms in the active set is close to the performance of the oracle arm up to four times $c(h,\delta_t)$.

Lemma 1. *On the h.p. event ξ_t, 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*

$$\overline{\mu}_i^h(N_{i,t}^{\pi_F}) \geq \mu_t^+(\pi_F) - 4c(h,\delta_t). \quad (5)$$

This result heavily relies on the non-increasing assumption of rotting bandit. In fact, for any arm i and any window h we have

$$\overline{\mu}_i^h(N_{i,t}^{\pi_F}) \geq \overline{\mu}_i^1(N_{i,t}^{\pi_F}) \geq \mu_i(N_{i,t}^{\pi_F}).$$

While the inequality above for i_t^* trivially satisfies Eq. 5, Lem. 1 is proved by integrating the possible errors introduced by the filter in selecting active arms due to the error of the empirical estimates.

Relating FEWA to the oracle policy. While Lem. 1 provides a link between the value of the arms returned by the filter and the oracle arm, $i_t^\star$ is defined according to the number of pulls obtained by FEWA up to t, which may significantly differ from the sequence of pulls of the oracle policy. 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. Let $h_{i,t} \triangleq \left|N_{i,t}^{\pi_F} - N_{i,t}^{\pi^\star}\right|$ be the absolute difference in the number of pulls between π_F and the optimal policy up to t. Since $\sum_i N_{i,t}^{\pi_F} = \sum_i N_{i,t}^{\pi^\star} = t$, we have that $\sum_{i\in OP} h_{i,t} = \sum_{i\in UP} h_{i,t}$ which means that there are as many total overpulls than underpulls. Let $j \in UP$ be an underpulled arm[5] with $N_{j,T}^{\pi_F} < N_{j,T}^{\pi^\star}$, then, for all $s \in \{0, \dots, h_{j,t}\}$, we have the inequality

$$\mu_T^+(\pi_F) = \max_{i\in\mathcal{K}} \mu_i(N_{i,T}^{\pi_F}) \geq \mu_j(N_{j,T}^{\pi_F} + s). \quad (6)$$

As a result, from Eq. 3 we have the regret upper bound

$$R_T(\pi_F) \leq \sum_{i\in OP} \sum_{h=0}^{h_{i,T}-1} \Big(\mu^+(\pi_F) - \mu_i(N_{i,T}^{\pi^\star} + h)\Big), \quad (7)$$

where the inequality is obtained by bounding $\mu_i(t') \leq \mu_T^+(\pi_F)$ in the first summation and then using $\sum_{i\in OP} h_{i,T} = \sum_{i\in UP} h_{i,T}$. While the previous expression shows that we can just 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 $i \in OP$ 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 $i \in OP$ in Eq. 7 by applying a corollary of Lem. 1.

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

$$\mu_t^+(\pi_F) - \overline{\mu}_i^{h_{i,t}}(N_{i,t}) \leq 4c(h_{i,t}, \delta_t). \quad (8)$$

4.2 Problem-Dependent Bounds

Since our setting generalizes the standard stochastic bandit setting, 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 Thm. 1 we actually proved that FEWA nearly matches the problem-independent regret $\widetilde{\mathcal{O}}(\sqrt{KT})$. We may wonder whether this is true for the *problem-dependent* regret as well.

Remark 1. *Consider a stationary stochastic bandit setting with expected rewards $\{\mu_i\}_i$ and $\mu_\star \triangleq \max_i \mu_i$. Corollary 1 guarantees that for $\delta_t \geq 1/(KT^\alpha)$,*

$$\mu_\star - \mu_i \leq 4c(h_{i,T}-1, \delta_t) = 4\sqrt{\frac{2\alpha\sigma^2\log(KT)}{h_{i,T}-1}}$$

$$\text{or equivalently, } h_{i,T} \leq 1 + \frac{32\alpha\sigma^2\log(KT)}{(\mu_\star - \mu_i)^2}. \quad (9)$$

Therefore, our algorithm matches the lower bound of Lai and Robbins (1985) up to a constant, thus showing that learning in the rotting bandit is never harder than in the stationary case. Moreover, this upper bound is at most α larger than the one for UCB1 (Auer et al., 2002a).[6] 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 a factor 4 in the confidence bound guarantee on the selected arm (vs 2 for UCB), which implies 4 times more overpulls than UCB (see Eq. 9). 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 better selection rule. The other source of suboptimality w.r.t. UCB is the use of larger confidence bands because of the higher number of estimators computed at each round (Kt^2 instead of Kt for UCB).

Remark 1 also reveals that Corollary 1 can be used to derive a general problem-dependent result in the rotting case. In particular, with Corollary 1 we upper-bound the maximum number of overpulls by a problem dependent quantity

$$h_{i,T}^+ \triangleq \max\left\{h \leq 1 + \frac{32\alpha\sigma^2\log(KT)}{\Delta_{i,h-1}^2}\right\}, \quad (10)$$

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

We then use Corollary 1 again to upper-bound the regret caused by $h_{i,T}^+$ overpulls for each arm, leading to Corollary 2 (see proof in App. D).

Corollary 2. *For $\delta_t \triangleq 1/(Kt^5)$ and $C_\alpha \triangleq 32\alpha\sigma^2$, the regret of FEWA is bounded as*

$$\mathbb{E}[R_T(\pi_F)] \leq \sum_{i\in\mathcal{K}} \left(\frac{C_5\log(KT)}{\Delta_{i,h_{i,T}^+-1}} + \sqrt{C_5\log(KT)} + L\right).$$

5 Numerical Simulations

The 2-arm setting. We report numerical simulations designed to provide insights on the difference between

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

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

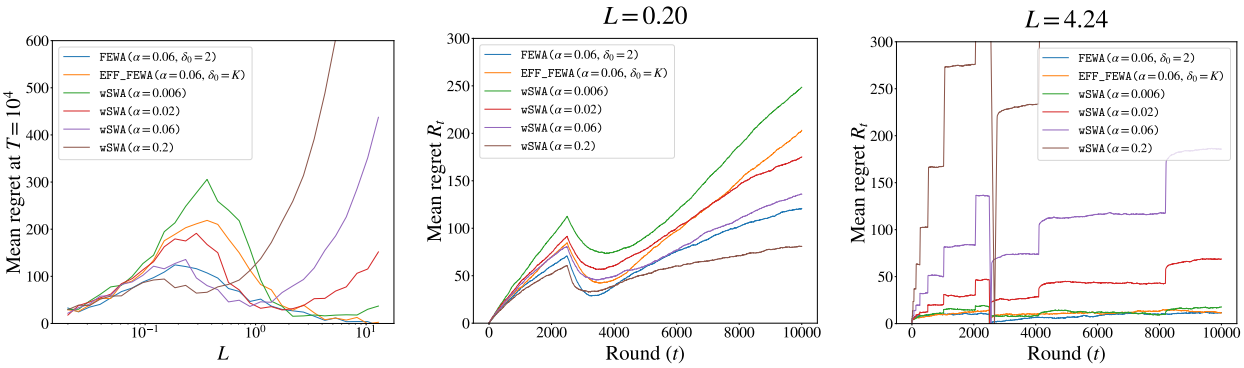

Figure 1: Comparison between FEWA and wSWA in the two-arm single-decrement case. **Left:** Regret at $T = 10,000$ for different values of L. **Middle-right:** Average regret during the game for $L = 0.2$ (i.e., worst case for FEWA) and $L = 4.24$ (case of $L \gg \sigma$).

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 Asm. 1 with a bounded decay L. It is important to notice that in this specific case, L also plays the role of defining the gap Δ between the arms, which is known to heavily impact the performance in the stochastic bandit and in the rotting bandit setting (see Cor.2). In particular, for any learning strategy, the gap between the two arms is always $\Delta = |\mu_1(n_1) - \mu_2(n_2)| = L/2$. We also recall that in the stochastic bandit case, the problem independent bound $\mathcal{O}(\sqrt{KT})$ is obtained by the worst-case choice of $\Delta = \sqrt{K/T}$. In the two-arm setting defined above, the optimal allocation is $N^\star_{1,T} = 3T/4$ and $N^\star_{2,T} = T/4$.

The algorithms. Both algorithms have a parameter α to tune. In wSWA, α is a multiplicative constant to tune the window. We try four different values of α, including the recommendation of Levine et al. (2017), $\alpha = 0.2$. In general, the smaller the α, the smaller the averaging window and the more reactive the algorithm is to large drops. Nonetheless, in stationary regimes, this may correspond to high variance and poor regret. On the other hand, a large value of α may reduce variance but increase the bias in case of rapidly rotting arms. Thm. 3.1 of Levine et al. (2017) reveals this trade-off in the regret bound of wSWA, which has a factor $(\alpha\mu_{\max} + \alpha^{-1/2})$, which $\mu_{\max}$ is the largest value of any arm. The best choice of α is then $\mu_{\max}^{-2/3}$, which reduces the previous constant to $\mu_{\max}^{1/3}$. In our experiment, $\mu_{\max} = L$ and we could expected that for any fixed α, wSWA may perform well in cases when $\alpha \approx \mu_{\max}^{-2/3}$, while the performance may significantly degrade when $\mu_{\max}$ is much larger.

In FEWA, α tunes the confidence $\delta_t = 1/(t^\alpha)$ used in $c(h, \delta_t)$. While our analysis suggests $\alpha = 5$, the analysis of confidence intervals, union bounds, and filtering algorithms is too conservative for a typical case. Therefore, we use more aggressive values $\alpha \in \{0.03, 0.06, 0.1\}$.

Experiments. In Fig. 1, we compare the performance of the two algorithms and their dependence on L. The first plot shows the regret at T for various values of L and different algorithms. The second and the third plots show the regret as a function of time 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 averaged 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_2 as reference, π_2 would have an increasing regret for the first $T/4$ rounds, which then would plateau from $T/4$ up to $3T/4$ as both π_1 and π_2 are pulling arm 1. Then from $3T/4$ to T, the regret of π_1 would reverse back to 0 since π_2 would keep selecting arm 1 getting a reward of 0, while π_1 transitions to pulling arm 2 with a reward of $L/2$.

Results. Fig. 2 shows that the performance of wSWA depends on the proper tuning of α w.r.t. $\mu_{\max} = L$, as predicted by Thm.3.1 of Levine et al. (2017). In fact, for small values of L, the best choice is $\alpha = 0.2$, while for larger values of L a smaller α is preferable. In particular, when L grows very large, the regret tends to

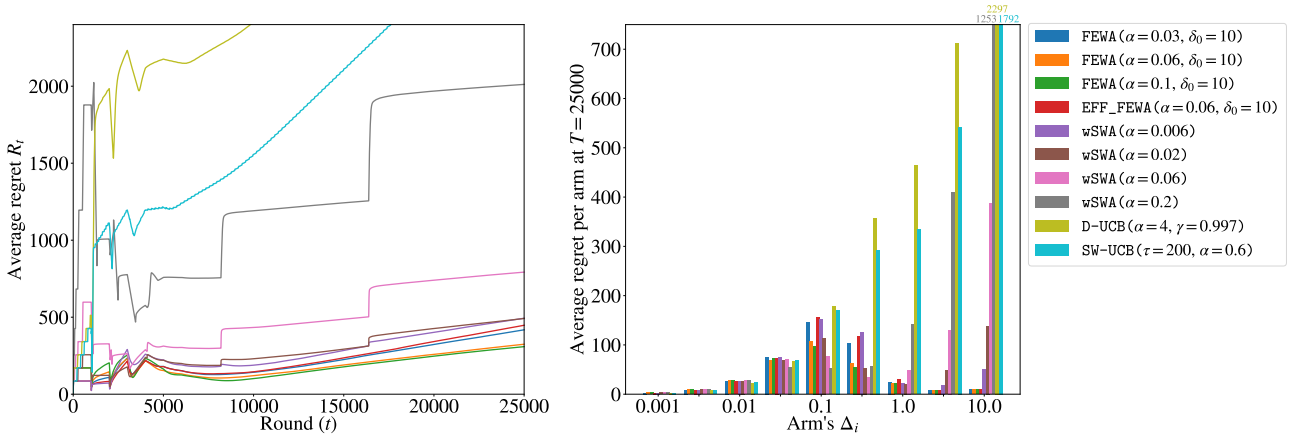

Figure 2: The setting with 10 arms. **Left :** Average regret during the game. **Right :** Average "regret per arm" at the end of the game.

grow linearly with L. On the other hand, FEWA seems much more robust to different values of L. Whenever T and σ are large compared to L, Thm. 1 suggests that the regret of FEWA is dominated by $\mathcal{O}(\sigma\sqrt{KT})$, while the term KL becomes more relevant for large value of the drop L. We also notice that as L defines the gap between the value of μ_1 and μ_2, the problem-independent bound is achieved for the worst-case choice of $L \sim 2\sqrt{K/T}$, when the regret of FEWA is indeed the largest. Fig. 1 middle and right confirm these findings for the extreme choice of the worst-case value of L and the regime where the drop is much larger than the noise level (i.e., where the term KL dominates the regret). 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 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.

The 10-arm setting. 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. In this setting, the regret can be written $R_T(\pi) = \sum_{i=1}^{9} h_{i,T}\Delta_i$. Hence, one could define the regret per arm:

$$R_T^i(\pi) \triangleq \Delta_i h_{i,T}.$$

In Figure 2, we compare the performance of different algorithms on this setting. For each algorithm, we retain only the best parameter (tested over a grid of parameter). The left plot shows the average regret as a function of time. The right plot shows the regret per arm (indexed by their Δ_i) at the end of the game.

Results. The 2-arm setting shows that wSWA has to be tuned to the single decrement size to be competitive while FEWA is always competitive. How the different algorithms can manage several decrement size in the same game? On the left figure, we see that FEWA outperform wSWA at the end of the game. We remark that the best tuning for wSWA corresponds to a rather small window which is good around L = 2 in the 2-arms settings. Similar result can be observed on the right figure : wSWA slightly outperforms FEWA for $\Delta_i = 0.3$ and $\Delta_i = 1$. But this single window is too large for $\Delta_i = 3.2$ and $\Delta_i = 10$. Indeed, we can see on the left figure that for the two last doubling tricks, the window is increased which leads to extra pulls of these two arms and ultimately to sharp regret increment. We also remark that EFF-FEWA is still penalized by arm with rather small Δ_i, for which the impact of the delay is more important.

We also add SW-UCB and D-UCB (Garivier and Moulines, 2011) with forgetting parameters tuned for this experimental setup. While the two algorithms are known benchmarks for non-stationary *restless* bandits, they are both heavily penalized on our *rested* bandits problem. Indeed, they keep exploring arms that have not been pulled for many rounds which is detrimental in our case as the arms stay constant when they are not pulled. Hence, there is no good choice of their forgetting parameters τ and γ 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.

6 Conclusion

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 $\widetilde{\mathcal{O}}(K^{1/3}T^{2/3})$. Thus our result shows that the rotting bandit scenario is not harder than the stochastic setting. Our technical analysis of FEWA hinges on the *adaptive* nature of the window size. The most interesting aspect of the proof technique 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.

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.

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.

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.

Romain Warlop, Alessandro Lazaric, and Jérémie Mary. Fighting boredom in recommender systems with linear reinforcement learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 1764–1773, 2018.

A Proof of core FEWA guarantees

Lemma 1. *On the h.p. event ξ_t, 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*

$$\overline{\mu}_i^h(N_{i,t}^{\pi_\mathrm{F}}) \geq \mu_t^+(\pi_\mathrm{F}) - 4c(h,\delta_t). \tag{5}$$

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 $\widehat{\mu}_{\max,t}^h$,

$$\overline{\mu}_i^h(N_{i,t}) \geq \widehat{\mu}_i^h(N_{i,t}) - c(h,\delta_t) \geq \widehat{\mu}_{\max,t}^h - 3c(h,\delta_t) \geq \max_{i\in\mathcal{K}_h} \overline{\mu}_i^h(N_{i,t}) - 4c(h,\delta_t), \tag{11}$$

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

$$\mu_t^+(\pi_\mathrm{F}) \triangleq \mu_{i_t^\star}(N_{i_t^\star,t}) \leq \mu_{i_t^\star}(N_{i_t^\star,t}-1) = \overline{\mu}_{i_t^\star}^1(N_{i_t^\star,t}) \leq \max_{i\in\mathcal{K}} \overline{\mu}_i^1(N_{i,t}) = \max_{i\in\mathcal{K}_1} \overline{\mu}_i^1(N_{i,t}). \tag{12}$$

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,

$$\max_{i\in\mathcal{K}_{h'}} \overline{\mu}_i^{h'}(N_{i,t}) = \overline{\mu}_{i_{\max}^{h'}}^{h'}(N_{i_{\max}^{h'},t}) \leq \overline{\mu}_{i_{\max}^{h'}}^{h'+1}(N_{i_{\max}^{h'},t}) \leq \max_{i\in\mathcal{K}_{h'+1}} \overline{\mu}_i^{h'+1}(N_{i,t}), \tag{13}$$

because $i_{\max}^{h'} \in \mathcal{K}_{h'+1}$. Gathering Equations 11, 12, and 13 leads to the claim of the lemma,

$$\overline{\mu}_i^h(N_{i,t}) \overset{(11)}{\geq} \max_{i\in\mathcal{K}_h} \overline{\mu}_i^h(N_{i,t}) - 4c(h,\delta_t) \overset{(13)}{\geq} \max_{i\in\mathcal{K}_1} \overline{\mu}_i^1(N_{i,t}) - 4c(h,\delta_t) \overset{(12)}{\geq} \mu_t^+(\pi_\mathrm{F}) - 4c(h,\delta_t).$$

□

Corollary 1. *Let $i \in$ 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 the h.p. event ξ_t, we have*

$$\mu_t^+(\pi_\mathrm{F}) - \overline{\mu}_i^{h_{i,t}}(N_{i,t}) \leq 4c(h_{i,t},\delta_t). \tag{8}$$

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

$$\overline{\mu}_i^h(N_{i,t}) \geq \mu_t^+(\pi_\mathrm{F}) - 4c(h_{i,t},\delta_t). \tag{14}$$

□

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,

$$R_T(\pi) \leq \sum_{i\in\mathrm{OP}} \sum_{t'=0}^{h^\pi_{i,T}-1} \left(\mu^+_T(\pi_\mathrm{F}) - \mu_i(N^{\pi^\star}_{i,T}+t')\right). \tag{15}$$

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

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

□

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_{π_F} defined in Lemma 2 is upper-bounded by*

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

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

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

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

$$\begin{aligned} A_{\pi_\text{F}} &\leq \sum_{i\in \text{OP}_\xi} \mu_T^+(\pi_\text{F}) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,T} - 1\right) + 4\left(h^{\xi}_{i,T}-1\right)c\left(h^{\xi}_{i,T}-1, \delta_{t_i}\right) && (16) \\ &\leq \sum_{i\in \text{OP}_\xi} \mu_T^+(\pi_\text{F}) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,T} - 1\right) + 4\left(h^{\xi}_{i,T}-1\right)c\left(h^{\xi}_{i,T}-1, \delta_{T}\right) && (17) \\ &\leq \sum_{i\in \text{OP}_\xi} \mu_T^+(\pi_\text{F}) - \mu_i\left(N^{\star}_{i,T} + h^{\xi}_{i,T} - 1\right) + 4\sqrt{2\alpha\sigma^2\left(h^{\xi}_{i,T}-1\right)\log_+\left(KT\delta_0^{-1/\alpha}\right)}, && (18) \end{aligned}$$

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 $\log_+(KT^\alpha) \leq \alpha\log_+(KT)$ for $\alpha > 1$. If $N^{\pi^\star}_{i,T} = 0$ and $h^{\xi}_{i,T} = 1$ then

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

since and $\mu^+(\pi_\text{F}) \leq L$ and $\mu_i(0) \geq 0$ 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

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

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 18 as

$$A_{\pi_{\mathrm{F}}} \leq \sum_{i \in \mathrm{OP}_\xi} \left(4\sqrt{2\alpha\sigma^2 \log_+ \left(KT\delta_0^{-1/\alpha}\right)} + 4\sqrt{2\alpha\sigma^2 \left(h_{i,T}^{\xi} - 1\right) \log_+ \left(KT\delta_0^{-1/\alpha}\right)} + L \right). \tag{19}$$

□

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*

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

C Minimax regret analysis of FEWA

Theorem 1. *For any rotting bandit scenario with means $\{\mu_i(n)\}_{i,n}$ satisfying Asm. 1 with bounded decay L and any time horizon T,* FEWA *run with $\alpha = 5$ and $\delta_t = 1/(Kt^5)$, suffers an expected regret*[7]

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

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 $A_{\pi_{\mathrm{F}}}$ 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

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

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., $h_{i,T} = T/(K-1)$,

$$\begin{aligned} A_{\pi_{\mathrm{F}}} &\leq (K-1)(4\sqrt{10\sigma^2 \log(KT)} + L) + 4\sqrt{10\sigma^2 \log(KT)} \sum_{i \in \mathrm{OP}} \sqrt{(h_{i,T} - 1)} \\ &\leq (K-1)(4\sqrt{10\sigma^2 \log(KT)} + L) + 4\sqrt{10\sigma^2 (K-1) T \log(KT)}. \end{aligned} \tag{20}$$

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

$$\mathbb{E}[B] \leq \frac{L\zeta(2)}{2} = \frac{L\pi^2}{12}. \tag{21}$$

Therefore, using Lemma 2 together with Equations 20 and 21, we bound the total expected regret as

$$\mathbb{E}[R_T(\pi_{\mathrm{F}})] \leq 4\sqrt{10\sigma^2 (K-1) T \log(KT)} + (K-1)(4\sqrt{10\sigma^2 \log(KT)} + L) + \frac{L\pi^2}{6}. \tag{22}$$

□

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

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

[7] 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

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

□

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,

$$A_{\pi_{\mathrm{F}}}\leq\sum_{i\in\mathrm{OP}_\xi}\left(4\sqrt{2\alpha\sigma^2\log(KT\delta_0^{-1/\alpha})}\Big(1+\sqrt{h^\xi_{i,T}-1}\Big)\right)+(K-1)L. \tag{23}$$

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

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

with $\mu^-_T(\pi^\star)\triangleq\min_{i\in\mathcal{K}}\mu_i\big(N^\star_{i,T}-1\big)$ 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 $\Delta_{i,h}\triangleq\mu^-_T(\pi^\star)-\overline{\mu}_i^h\big(N^\star_{i,t}+h\big)$ 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 $h^\xi_{i,T}$,

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

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,

$$h^\xi_{i,T}\leq h^+_{i,T}\triangleq\max\left\{h\leq T\Big|\ h\leq 1+\frac{32\alpha\sigma^2\log_+\Big(KT\delta_0^{-1/\alpha}\Big)}{\Delta^2_{i,h-1}}\right\}. \tag{24}$$

Since the square-root function is increasing, we can upper-bound Equation 18 by replacing $h^{\xi}_{i,T}$ by its upper bound $h^{+}_{i,T}$ to get

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

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

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

□

Corollary 2. *For $\delta_t \triangleq 1/(Kt^5)$ and $C_\alpha \triangleq 32\alpha\sigma^2$, the regret of* FEWA *is bounded as*

$$\mathbb{E}[R_T(\pi_{\mathrm{F}})] \leq \sum_{i\in\mathcal{K}} \left(\frac{C_5 \log(KT)}{\Delta_{i,h^{+}_{i,T}-1}} + \sqrt{C_5 \log(KT)} + L \right).$$

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

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

□

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

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

Proof. We consider the event $\cup_{t\leq T}\xi_t$ 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 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 $j \leq \log_2(N^{\pi_{\mathrm{EF}}}_{i,t})$, 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.

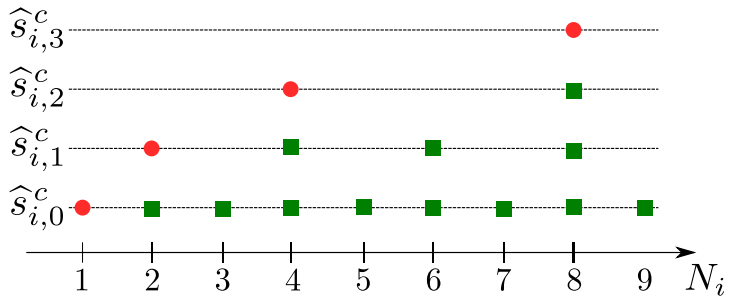

Figure 3: Illustration of the functioning of EFF-FEWA. The red circles denotes the number of pulls of arm i at which a new estimate $\widehat{s}^c_{i,j}$ is created corresponding to a window $h = 2^j$, while the green boxes indicate the number of pulls for which $\widehat{s}^c_{i,j}$ is updated with the last 2^j samples.

Algorithm 3 EFF-FEWA

Input: $\mathcal{K}$, δ_0, α

pull each arm once, collect reward, and initialize $N_{i,K} \leftarrow 1$
for $t \leftarrow K+1, K+2, \dots$ **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

1: $N_{i(t),t} \leftarrow N_{i(t),t-1} + 1$
2: $R_i^{\text{total}} \leftarrow R_i^{\text{total}} + r$ {*keep track of total reward*}
3: **if** $\exists j$ such that $N_{i,t} = 2^j$ **then**
4: $\quad \widehat{s}^c_{i,j} \leftarrow R_i^{\text{total}}/N_{i,t}$ {*initialize new statistics*}
5: $\quad \widehat{s}^p_{i,j} \leftarrow 0$
6: $\quad n_{i,j} \leftarrow 0$
7: **end if**
8: **for** $j \leftarrow 0 \dots \log_2(N_{i,t})$ **do**
9: $\quad n_{i,j} \leftarrow n_i + 1$
10: $\quad \widehat{s}^p_{i,j} \leftarrow \widehat{s}^p_{i,j} + r$
11: $\quad$ **if** $n_{i,j} = 2^j$ **then**
12: $\quad\quad \widehat{s}^c_{i,j} \leftarrow \widehat{s}^p_{i,j}/2^j$
13: $\quad\quad n_{i,j} \leftarrow 0$
14: $\quad\quad \widehat{s}^p_{i,j} \leftarrow 0$
15: $\quad$ **end if**
16: **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,*

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

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

Corollary 5. *Let $i \in \text{OP}$ be an arm overpulled by* EFF-FEWA *at round t and $h^{\pi_{\text{EF}}}_{i,t} \triangleq N^{\pi_{\text{EF}}}_{i,t} - N^{\pi^\star}_{i,t} \geq 1$ 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*

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

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 $2^f \leq h^{\pi_{\text{EF}}}_{i,t} < 2^{f+1}$. Note that for $h^{\pi_{\text{EF}}}_{i,t} = 1$, then EFF-FEWA has

the same guarantee as FEWA since the first filter is always up to date. Then for $h_{i,t}^{\pi_{\mathrm{EF}}} \geq 2$,

$$\overline{\mu}_i^{1,h_{i,t}^{\pi_{\mathrm{EF}}}}(N_{i,t}) \geq \overline{\mu}_i^{1,2^f-1}(N_{i,t}) = \frac{\sum_{j=1}^{f} 2^{j-1}\overline{\mu}_i^{2^{j-1},2^j-1}}{2^f-1} \tag{25}$$

$$\geq \mu_t^+(\pi_{\mathrm{EF}}) - \frac{4\sum_{j=1}^{f} 2^{j-1}c(2^{j-1},\delta)}{2^f-1} = \mu_t^+(\pi_{\mathrm{EF}}) - 4c(1,\delta_t)\frac{\sum_{j=1}^{f}\sqrt{2}^{j-1}}{2^f-1} \tag{26}$$

$$= \mu_t^+(\pi_{\mathrm{EF}}) - 4c(1,\delta_t)\frac{\sqrt{2}^f-1}{(2^f-1)(\sqrt{2}-1)} \geq \mu_t^+(\pi_{\mathrm{EF}}) - 4c(1,\delta_t)\frac{1}{\sqrt{2}^f(\sqrt{2}-1)} \tag{27}$$

$$= \mu_t^+(\pi_{\mathrm{EF}}) - \frac{4\sqrt{2}}{\sqrt{2}-1}c\big(2^{f+1},\delta_t\big) \geq \mu_t^+(\pi_{\mathrm{EF}}) - \frac{4\sqrt{2}}{\sqrt{2}-1}c\big(h_{i,t}^{\pi_{\mathrm{EF}}},\delta_t\big), \tag{28}$$

where Equation 25 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 26 uses Lemma 6 and make the dependence of $c(2^{j-1},\delta)$ on j explicit. Next, Equations 27 and 28 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 10.

F Numerical Simulations: Stochastic Bandit

In Figure 4 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.

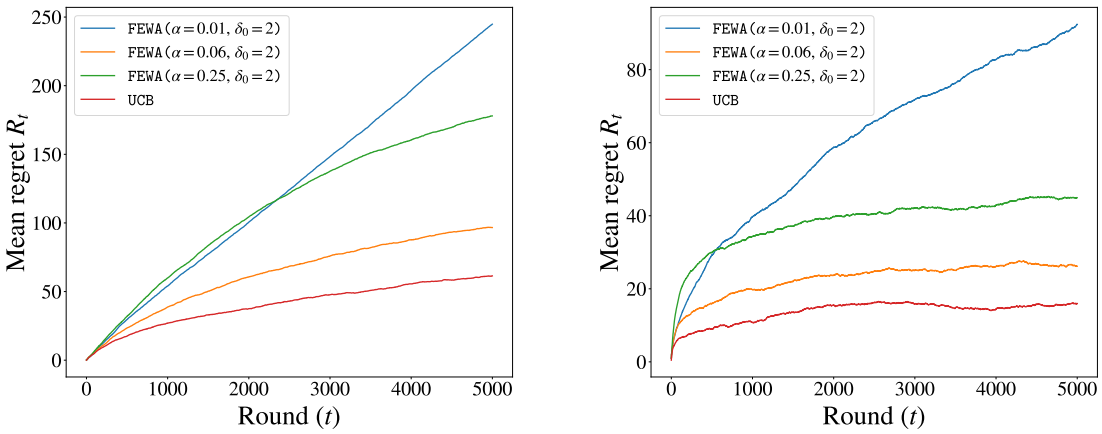

Figure 4: Comparing UCB1 and FEWA with $\Delta = 0.14$ and $\Delta = 1$.