ROTTING BANDITS ARE NOT HARDER THAN STOCHASTIC ONES

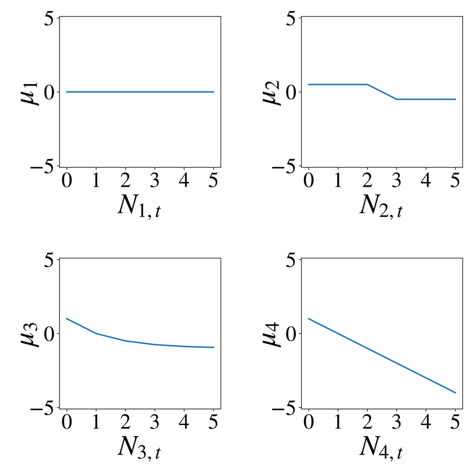
J. SEZNEC, A. LOCATELLI, A. CARPENTIER, A. LAZARIC, M. VALKO
julien.seznec@lelivrescolaire.fr

Setup

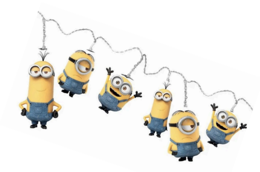
BANDITS:

At each round t:

- SELECT an action i
- RECEIVE noisy reward



ROTTING HYPOTHESIS:



- Each time we pull an arm, reward decay
- Maximum decay between two pulls: L

GOAL: Maximimize cumulative reward $\sum_{i\in\mathcal{K}} \sum_{n=0}^{N^{\pi}_{i,T}-1} \mu_i(n)$

APPLICATIONS: education, economics …

What's new?

FEWA algorithm

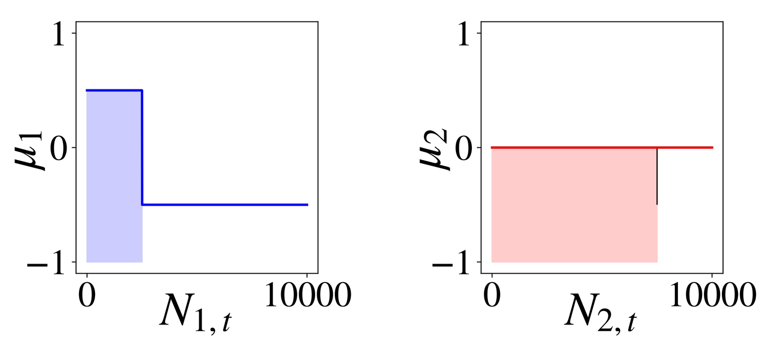
★ Minimax optimal. $T^{2/3} \to T^{1/2}$

★ Solves an open problem

★ First problem dependent guarantee - recovers bandits

★ Adaptive to L - does not need it as an input

Prior work



- Heidari et al. (2016)
  - **Optimal oracle - knows** μ_i
    → *Select the arm with largest next reward*
    Define regret against this optimal oracle

$$R_T(\pi) = \sum_{i\in \mathrm{UP}} \sum_{s=N^{\pi}_{i,T}+1}^{N^{\star}_{i,T}} \mu_i(s) - \sum_{i\in \mathrm{OP}} \sum_{s=N^{\star}_{i,T}+1}^{N^{\pi}_{i,T}} \mu_i(s)$$

  - **No noise - but unknown** μ_i
    → *Select the arm with largest last reward*
    → Is minimax optimal.

- Levine et al. (2017)
  - Noisy reward *sliding window average:*
    → *Select the arm with largest average of the* ***h*** *last reward sample* $\mathbb{E}\left[R_T(\mathrm{G})\right] \le KL$
  - Optimize ***h*** for the bias-variance trade-off

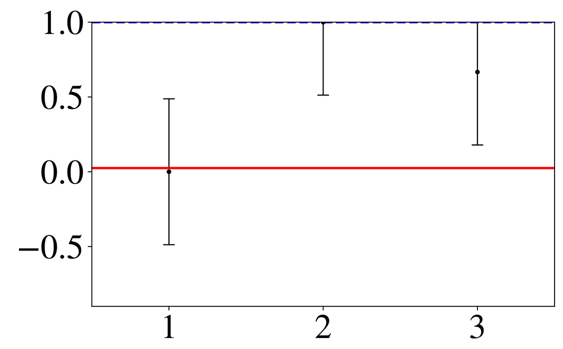
$$\mathbb{E}\left[R_T\left(\pi_{\mathrm{wSWA}}\right)\right] = \tilde{O}\left(L^{1/3}\sigma^{2/3}K^{1/3}T^{2/3}\right)$$

Algorithm

FEWA: Filtering on Expanding Window Average

Input: K, σ, α

1: **for** $t \leftarrow K+1, K+2, \dots$ **do**
2: $\quad \delta_t \leftarrow \frac{1}{Kt^{\alpha}}$
3: $\quad h \leftarrow 1$
4: $\quad \mathcal{K}_1 \leftarrow \mathcal{K}$
5: $\quad$ **do**
6: $\qquad \mathcal{K}_{h+1} \leftarrow \{i \in \mathcal{K}_h \,|\, \widehat{\mu}^h_i(N_{i,t}) \ge \max_{j\in\mathcal{K}} \widehat{\mu}^h_j(N_{j,t}) - 2c(h,\delta_t)\}$
7: $\qquad h \leftarrow h+1$
8: $\quad$ **while** $h \le \min_{i\in\mathcal{K}_h} N_{i,t}$
9: $\quad$ SELECT : $\{i \in \mathcal{K}_h \,|\, h > N_{i,t}\}$
10: **end for**



How does it work?

★ Method: Filter the set of arms

★ Based on: Expanding size of arms history (newest samples first)

★ Statistical tool: New way of using Hoeffding bound to both

★ Select relevant data history

★ Select arm maximizing exploration/exploitation tradeoff

Lemma

w.p. $1-\delta_t,\ \forall h \le N_{i,t},\ \bar{\mu}^h_i(N_{i,t}) \ge \max_{j\in\mathcal{K}} \mu_j(N_{j,t}) - 4c(h,\delta_t)$

Guarantees

WORST-CASE BOUND

$$\mathbb{E}\left[R_T\left(\pi_{\mathrm{F}}\right)\right] \le C\sigma\sqrt{KT}\log(KT) + KL$$

PROBLEM-DEPENDENT BOUND

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

$\Delta_{i,h}$ Difference between the worst reward pulled by the optimal policy and the average of the h first overpulls of arm i.

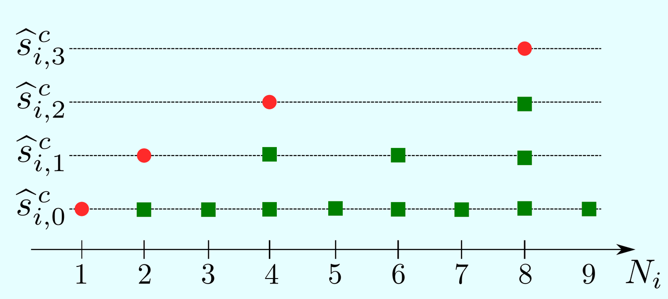
$h^+_{i,T}$ High-probability upper bound on the number of overpulls for FEWA.

Computational complexity

FEWA has an O(t) time and space complexity:

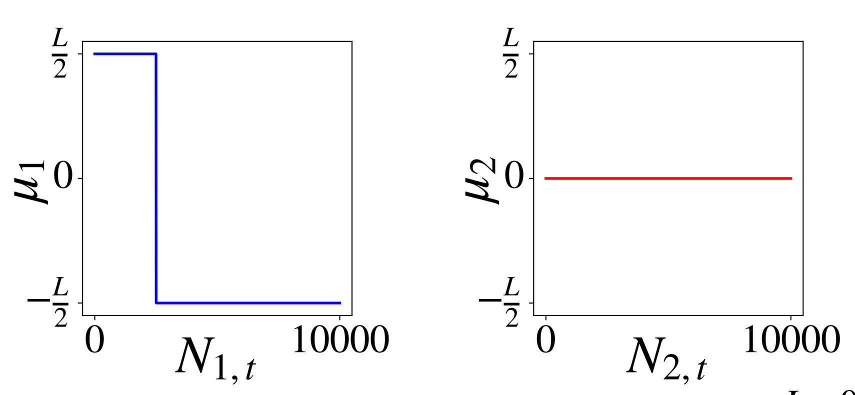
1. Perform log(t) filters: $h \leftarrow 2h$

2. Keep log(t) statistics:

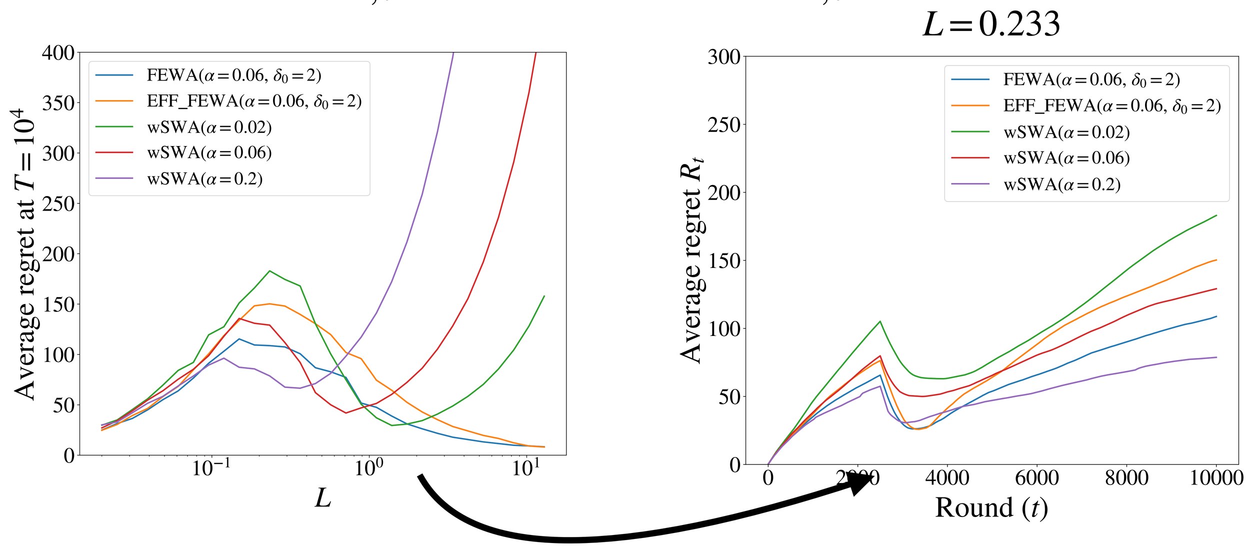


Experiments

2 arms - Minimal single drop experiment

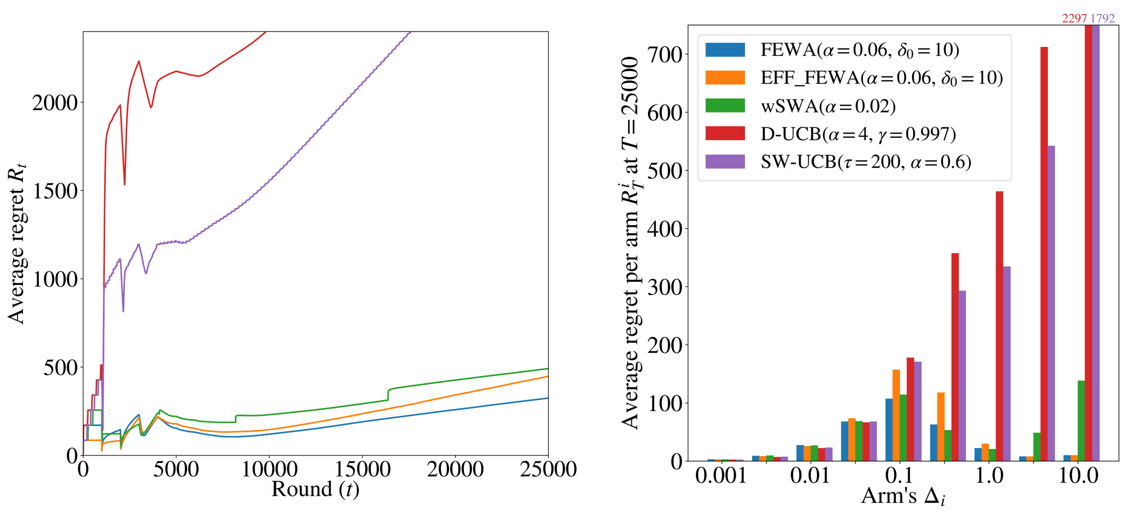


$L = 0.233$

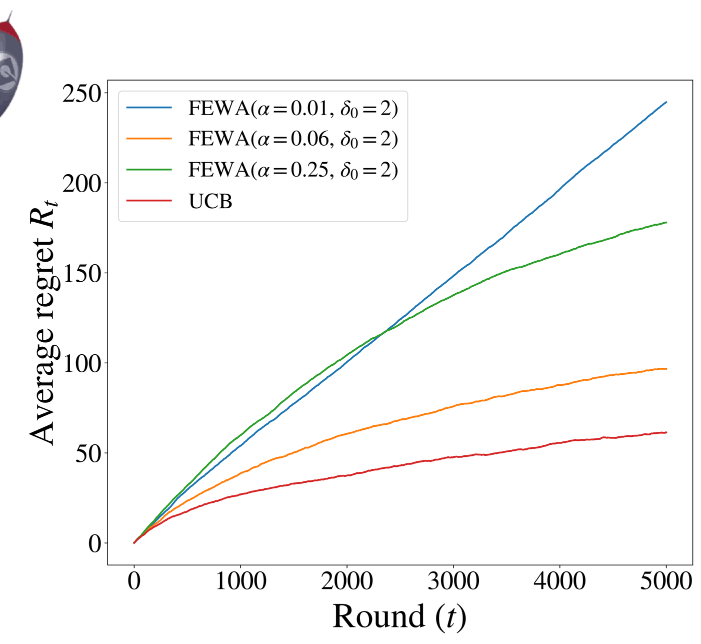


10 arms - Adapting to multiple decays

- 1 constant arm
- 9 arms with abrupt decay at 1000 pulls
- Geometric sequence of decays: 0.002 → 20



Competing against UCB1



1) Filtering policy < UCB index policy

2) More possible events → Looser CB