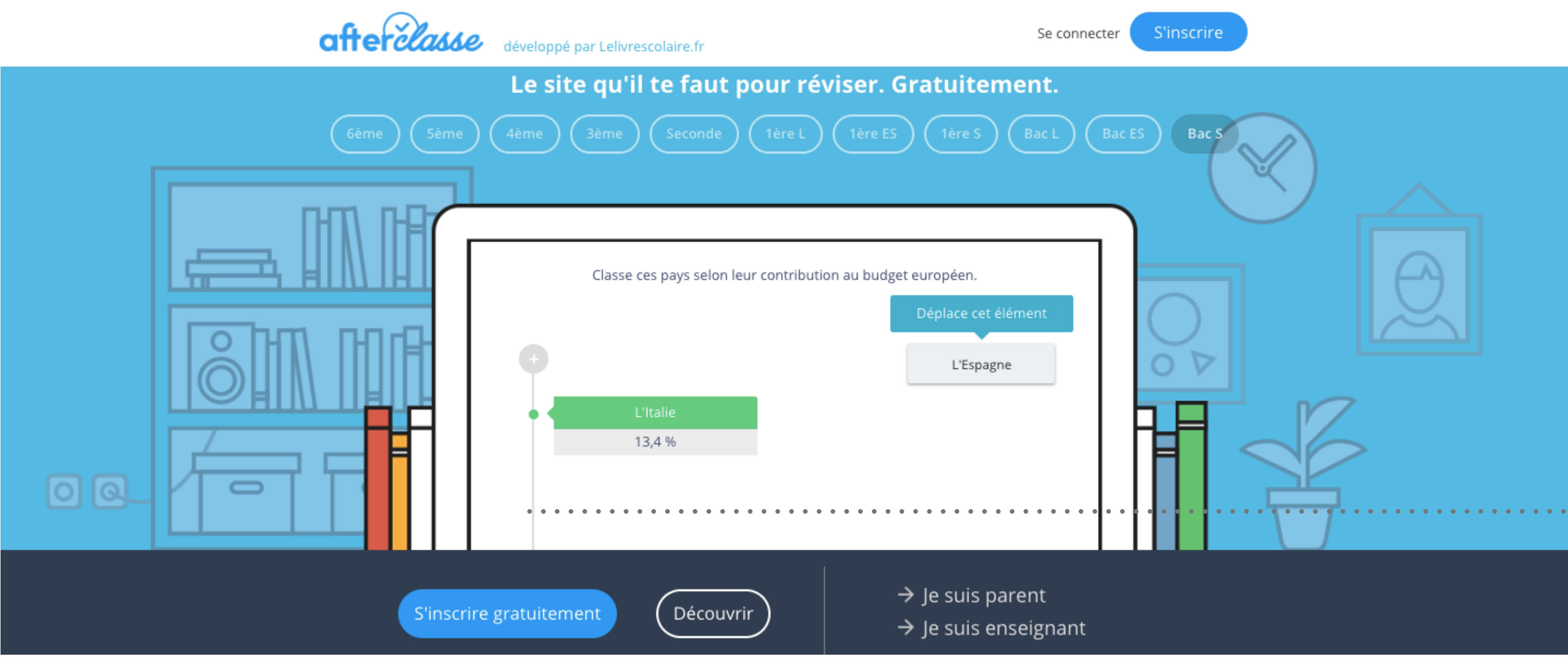


HOW DIFFICULT ARE ROTTING BANDITS?

lelivrescolaire.fr

https://www.afterclasse.fr

Michal Valko

with J. Seznec, A. Locatelli, A. Carpentier, A. Lazaric

WHEN BANDITS GO ROTTING …

CHAPITRE 1

L'origine des séismes et des éruptions volcaniques

CHAPITRE 2

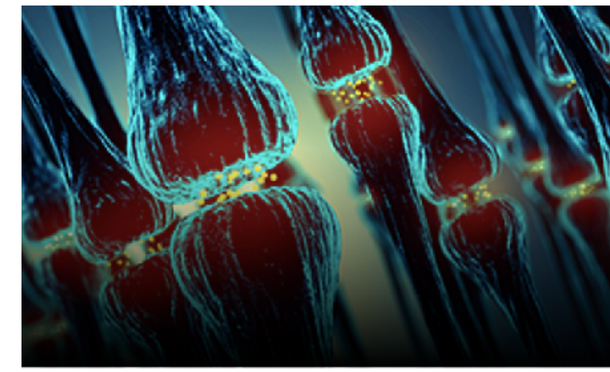
Les changements climatiques actuels et leurs conséquences

afterclasse

2 days before the national exam

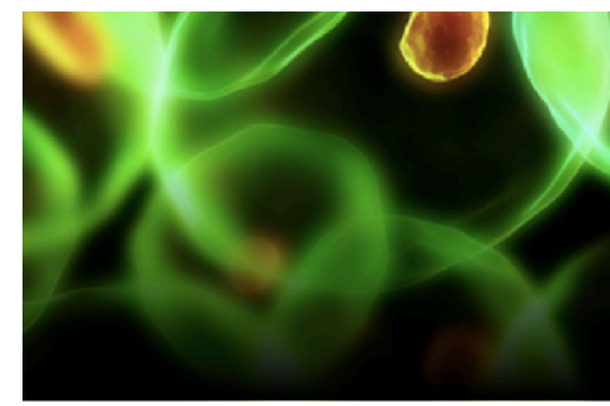
CHAPITRE 3

Les impacts des activités humaines sur l'environnement



CHAPITRE 8

Le fonctionnement du système nerveux



CHAPITRE 4

La nutrition à l'échelle cellulaire

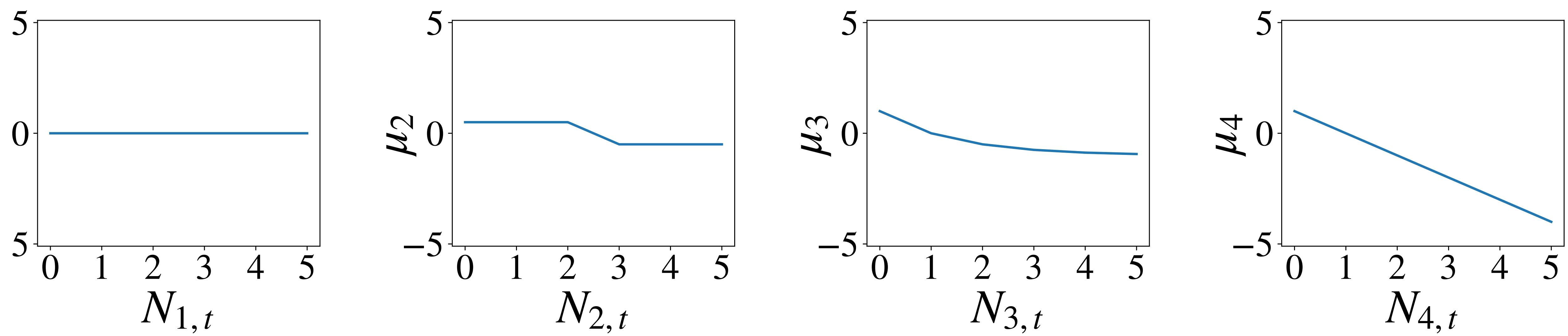
RESTED ROTTING BANDITS ARE …

Stochastic bandits …

- *K arms*
- *At each round t, agent pulls arm i and receives a noisy reward $r_t \leftarrow \mu_i + \varepsilon_t$*
  (ε_t i.i.d. ; σ-sub-gaussian)
- ***Maximize cumulative reward:*** $\mathbb{E}\left[\sum_{t \le T} r_t\right]$

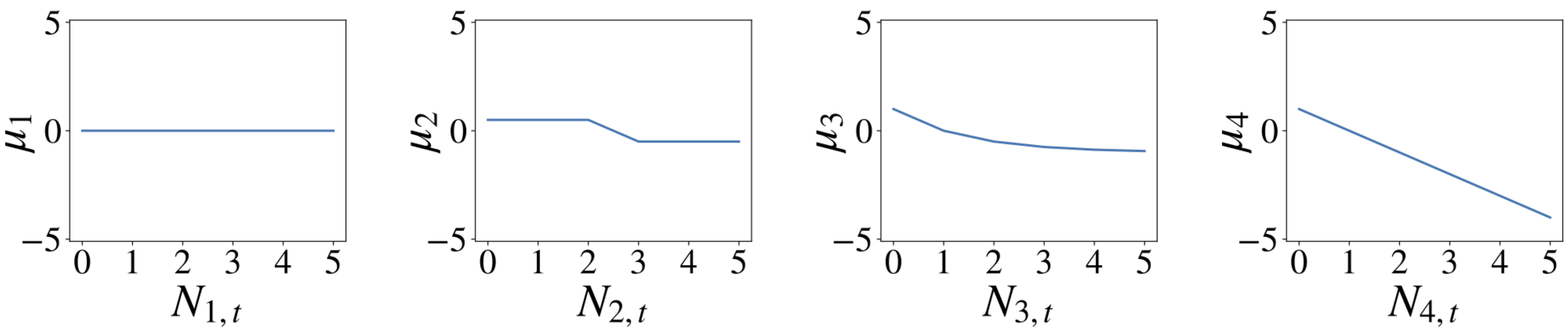
… with rotting arms

- $\{\mu_i\}$ are *non-increasing* functions of $N_{i,t}$ the *number of pulls of arm i* at time t
- $L \triangleq \max_{i \in K} \max_{n \le T} \mu_i(n) - \mu_i(n+1)$



BACK TO THE EXAMPLE

Bandit	Platform
K arms	K topics that can be revised, every time a topic is selected, a question is generated
Maximize cum. reward	maximize number of questions that students do not master
Observations r_t	$r_t = 0$ if answer is correct, $r_t = 1$ is answer is wrong
Rotting rewards	the student acquires knowledge over time



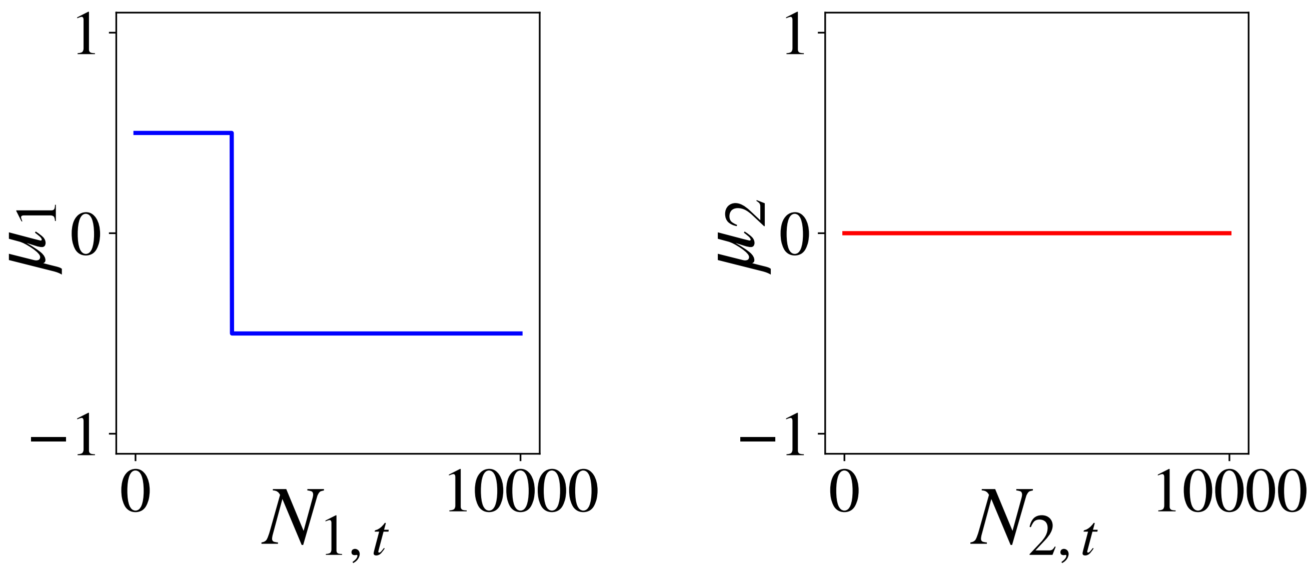
NOISELESS

$$\varepsilon = 0$$

OPTIMAL ORACLE POLICY [HEIDARI, 2016]

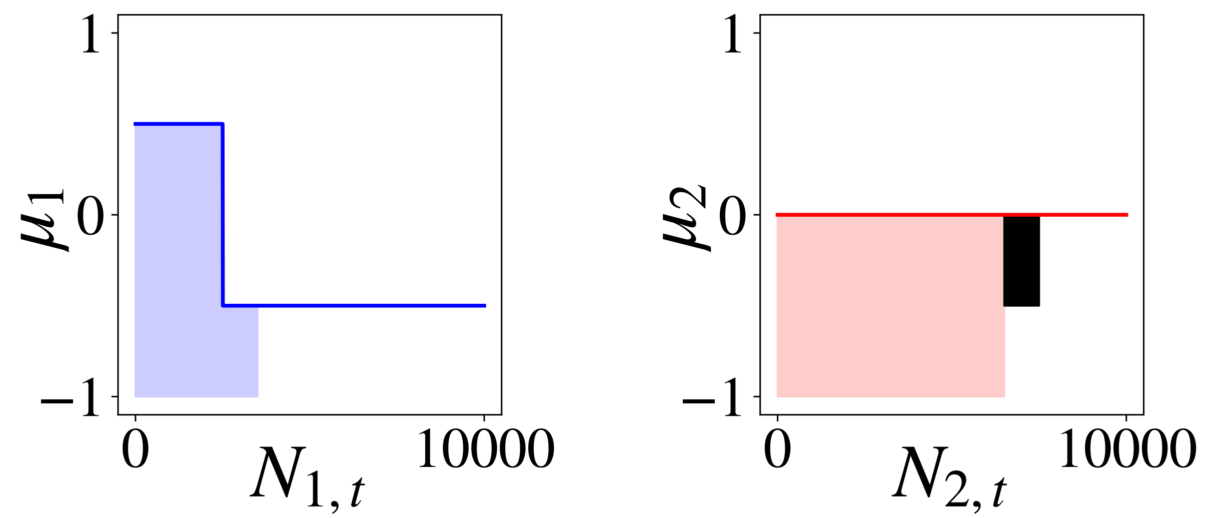
Algorithm 1 $\mathcal{A}_0$ (Heidari et al., 2016)

for $t \leftarrow 1, 2, \ldots$ **do**
  SELECT : $\arg\max_{i \in \mathcal{K}} \mu_i(N_{i,t})$
end for



OPTIMAL ORACLE POLICY [HEIDARI, 2016]

- $N^\star_{i,T}$ optimal num. of pulls by T
- UP $= \{i : N^\pi_{i,T} < N^\star_{i,T}\}$
- OP $= \{i : N^\pi_{i,T} > N^\star_{i,T}\}$



$$R_T(\pi) = \sum_{t=1}^{T} \mu_{i^\star(t)}(N^\star_{i^\star(t),t}) - \sum_{t=1}^{T} \mu_{i(t)}(N_{i(t),t})$$

$$\stackrel{*}{=} \sum_{i \in \text{UP}} \sum_{s=N^\pi_{i,T}+1}^{N^\star_{i,T}} \mu_i(s) - \sum_{i \in \text{OP}} \sum_{s=N^\star_{i,T}+1}^{N^\pi_{i,T}} \mu_i(s)$$

*order *does not* matter!

DETERMINISTIC CASE ($\varepsilon = 0$) [HEIDARI, 2016]

- Greedy *oracle* policy

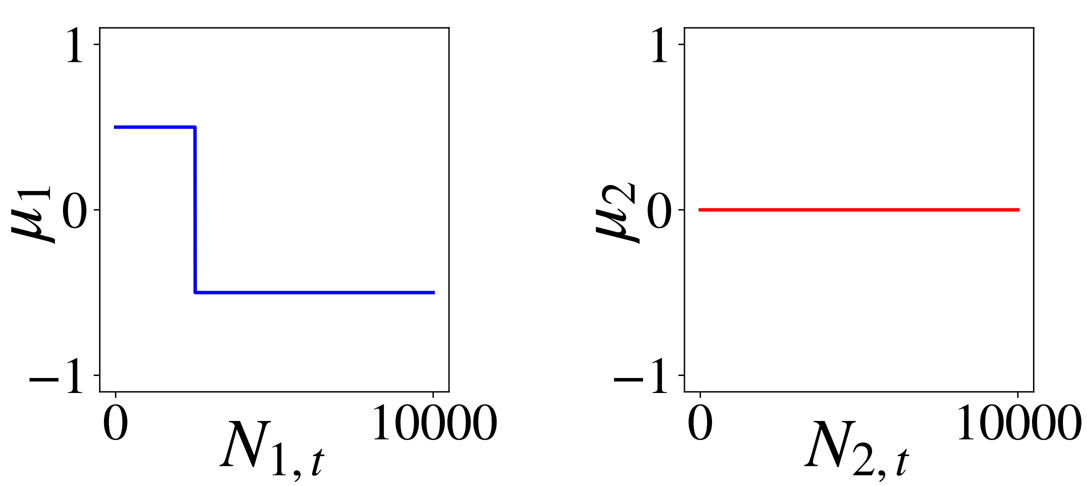
$$i^\star(t) = \arg\max_i \mu_i(N_{i,t})$$

- *Greedy policy*: select the arm with largest *last known* value

$$i(t) = \arg\max_i \mu_i(N_{i,t} - 1)$$

$$R_T(\pi) \leq K L$$

$\Rightarrow$ We pay for regret only *once per arm*



Algorithm 2 $\mathcal{A}_2$ (Heidari et al., 2016)

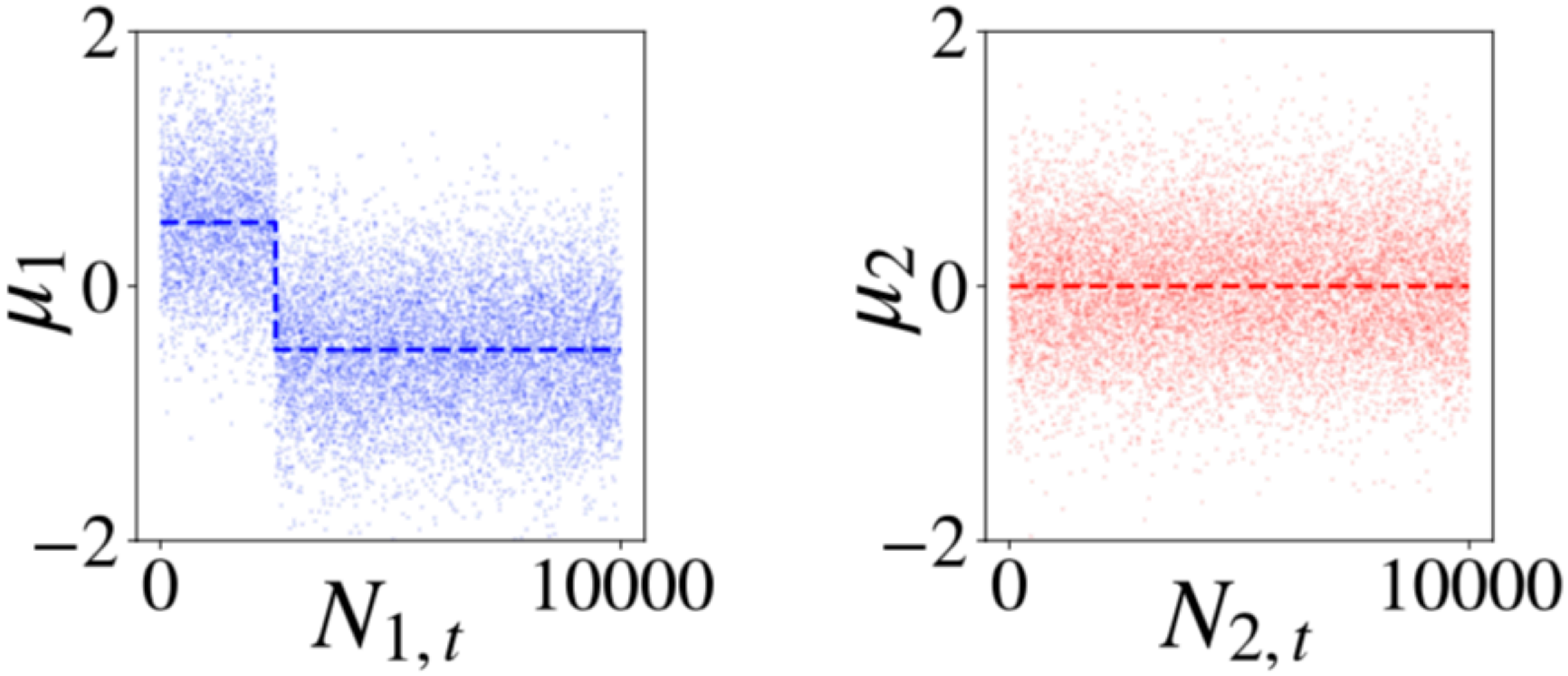
for $t \leftarrow K+1, K+2, \ldots$ **do**
    SELECT : $\arg\max_{i \in \mathcal{K}} \mu_i(N_{i,t} - 1)$
end for

NOISE

Noisy case [Levine et al., 2017]

Sliding-window Average of h most recent observations

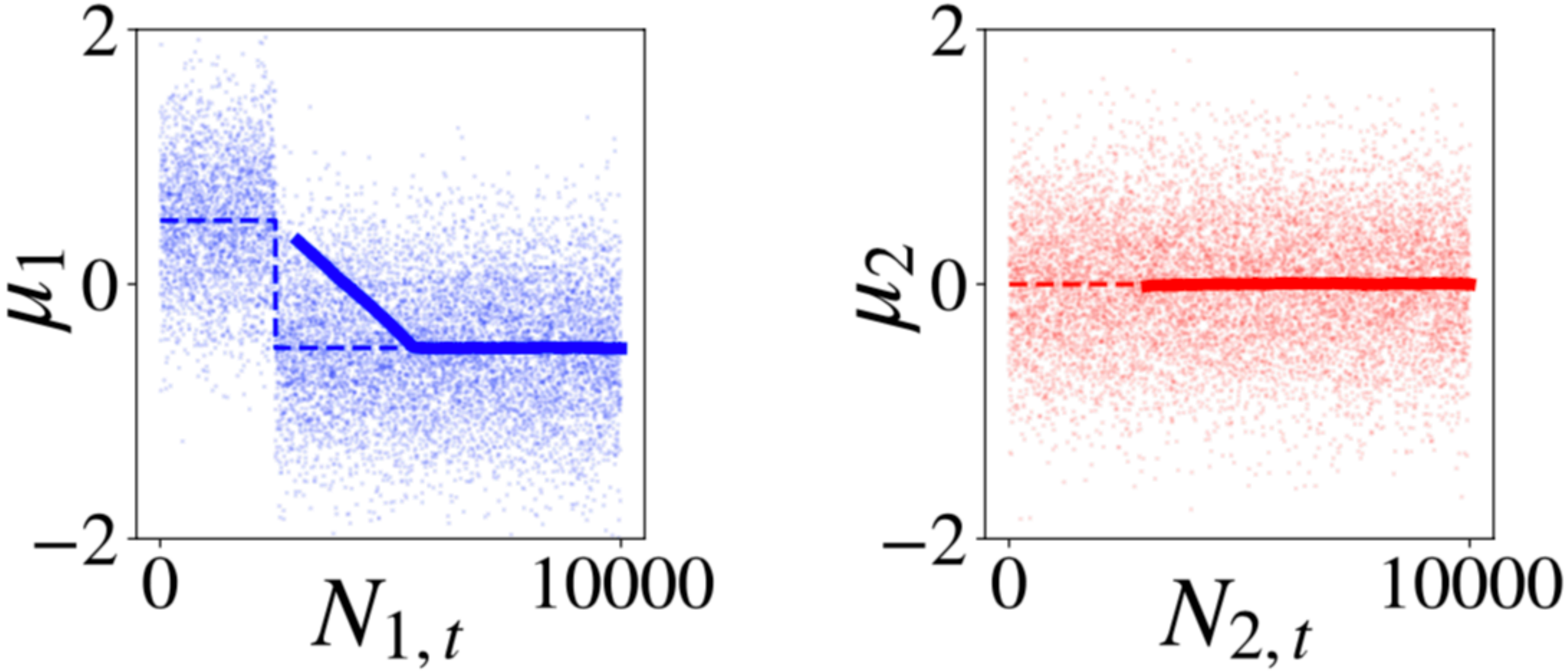
$$\widehat{\mu}_i^h(N_{i,t}) = \frac{1}{h}\sum_{j=1}^{h} r_i(N_{i,t}-j)$$



Noisy case [Levine et al., 2017]

Sliding-window Average of h most recent observations

$$\widehat{\mu}_i^h(N_{i,t}) = \frac{1}{h}\sum_{j=1}^{h} r_i(N_{i,t} - j)$$

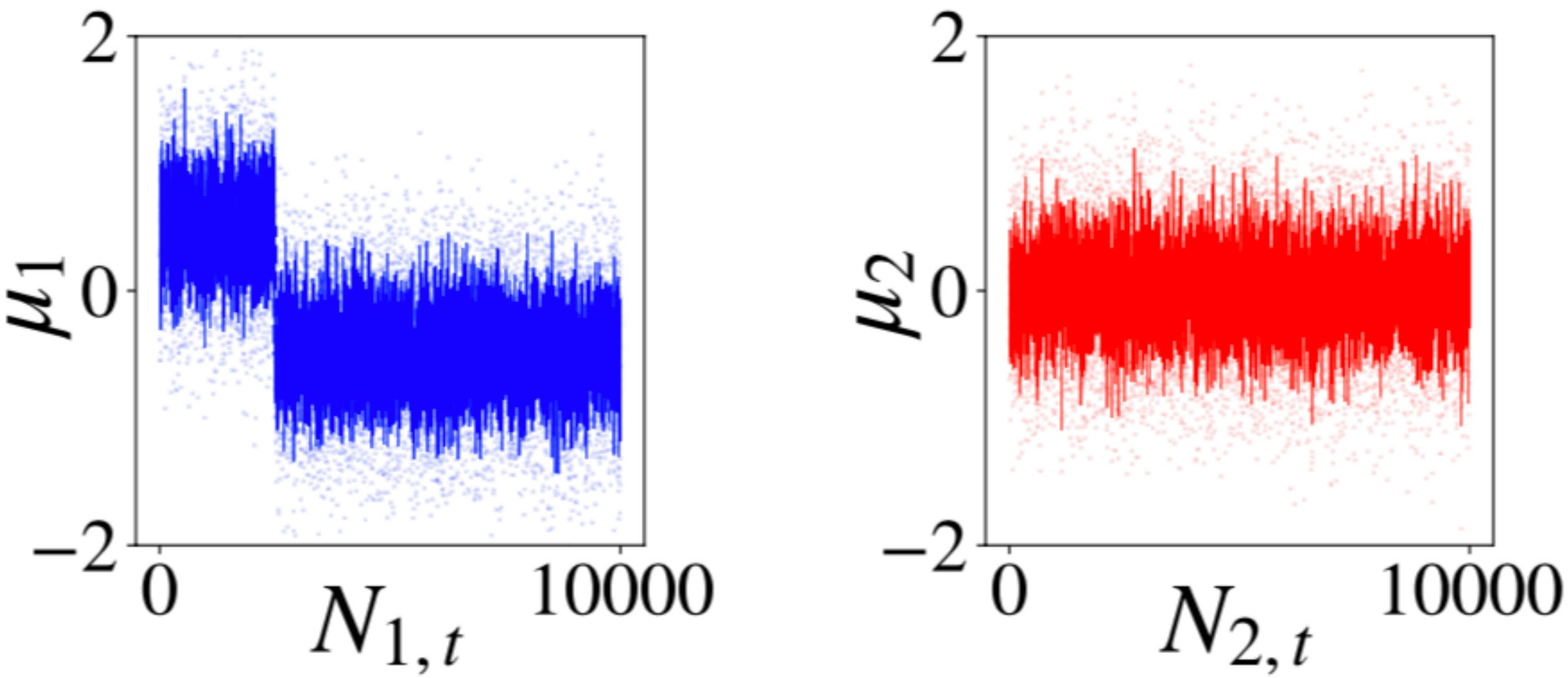


$h = 3000$ (*high* bias, *low* variance)

Noisy case [Levine et al., 2017]

Sliding-window Average of h most recent observations

$$\widehat{\mu}_i^h(N_{i,t}) = \frac{1}{h}\sum_{j=1}^{h} r_i(N_{i,t} - j)$$

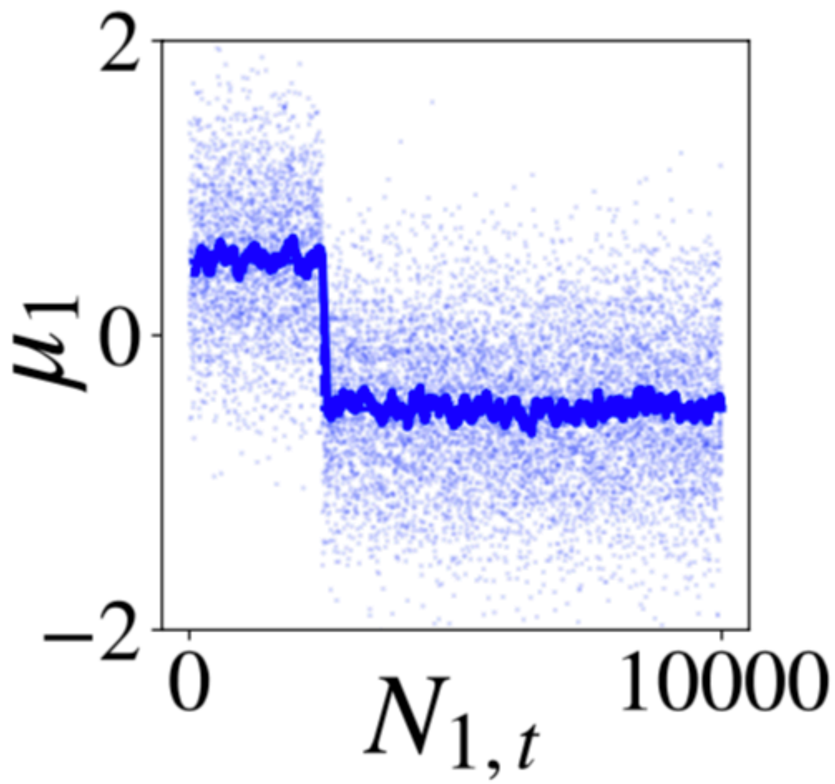


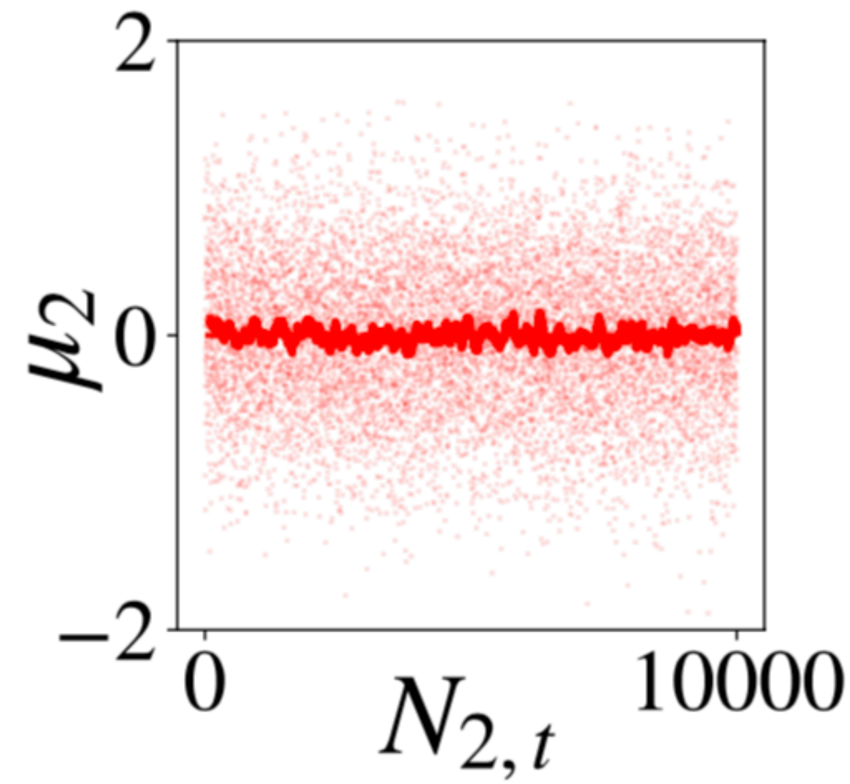
$h = 3$ (*low* bias, *high* variance)

Noisy case [Levine et al., 2017]

Sliding-window Average of h most recent observations

$$\widehat{\mu}_i^h(N_{i,t}) = \frac{1}{h}\sum_{j=1}^{h} r_i(N_{i,t} - j)$$





$h = 100$ (*ok* bias, *ok* variance)

wSWA [Levine et al., 2017]

shie@ee.technion.ac.il

THE FAILURE OF wSWA

Won't we benefit from a data-adaptive window ?

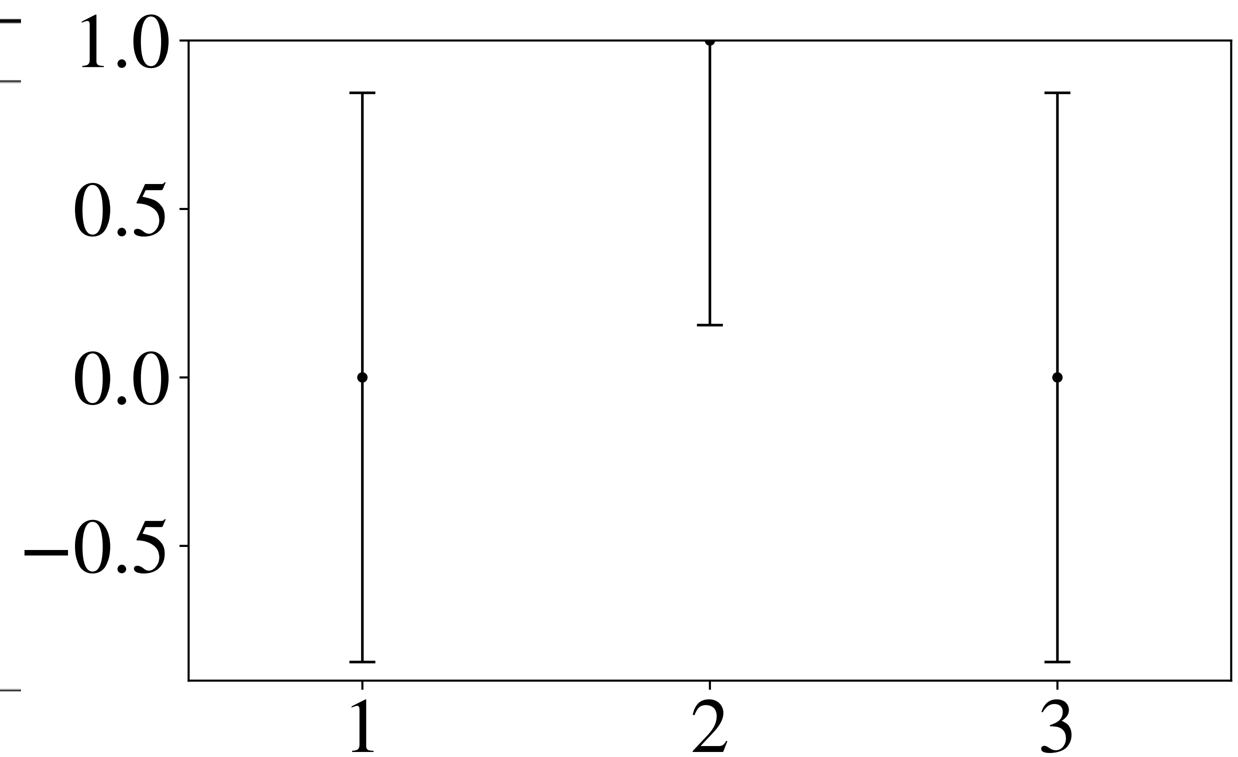
Sample		last
Arm 1	0	0
Arm 2	1	1
Arm 3	1	0

FILTERING ON EXPANDING WINDOW AVERAGE (THE MIDNIGHT FEWA)

Algorithm 4 FEWA

Input: K, σ, α

for $t \leftarrow K+1, K+2, \ldots$ **do**
  $\delta_t \leftarrow \frac{1}{Kt^\alpha}$
  $h \leftarrow 1$
  $\mathcal{K}_1 \leftarrow \mathcal{K}$
  do
    $\mathcal{K}_{h+1} \leftarrow \left\{i \in \mathcal{K}_h \,|\, \widehat{\mu}_i^h(N_{i,t}) \geq \max_{j\in\mathcal{K}} \widehat{\mu}_j^h(N_{j,t}) - 2c(h,\delta_t)\right\}$
    $h \leftarrow h+1$
  while $h \leq \min_{i\in\mathcal{K}_h} N_{i,t}$
  SELECT : $\{i \in \mathcal{K}_h \,|\, h > N_{i,t}\}$
end for



Sample	old																		last
Arm 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
Arm 2	1	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	1	1
Arm 3	X	X	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	0

UPPER BOUNDS

No L!

Same as for the deterministic case

Worst-case upper bound

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

Comparison w/ wSWA

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

Problem-dependent upper bound

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

Comparison w/ wSWA

Pure worst-case strategy

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

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

UPPER BOUNDS

Worst-case upper bound

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

Comparison w/ wSWA

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

Problem-dependent upper bound

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

Comparison w/ wSWA

Pure worst-case strategy

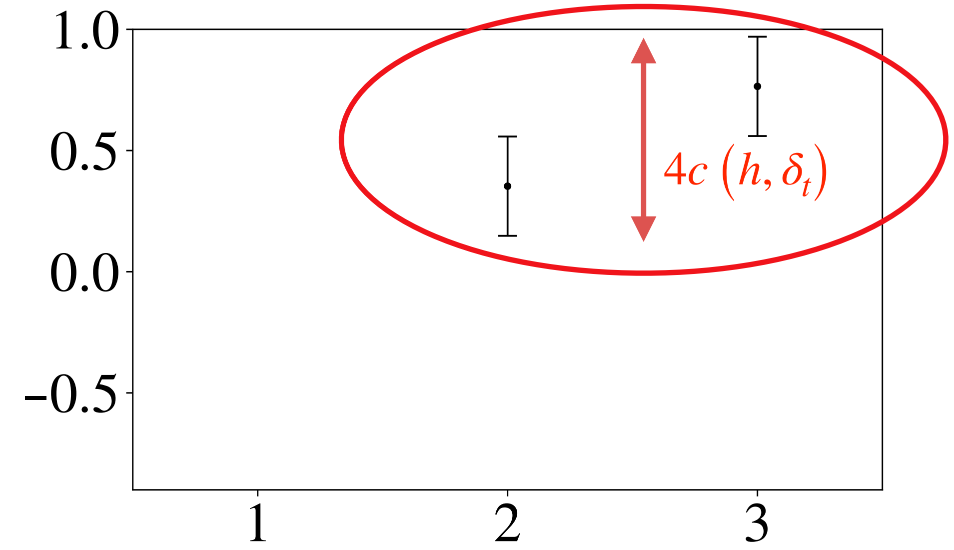
$\Delta_{i,h} = \Delta_i$ **on a stationary bandit problem**

$\Delta_{i,h^+_{i,T}-1}$ **is a problem-dependent quantity**

Proof sketch (for the instance-independent bound)

Lemma: ***if you pass, you can't be too bad***

$$\overline{\mu}_i^h(N_{i,t}) \geq \max_{i \in \mathcal{K}} \mu_i(N_{i,t}) - 4c(h, \delta_t)$$



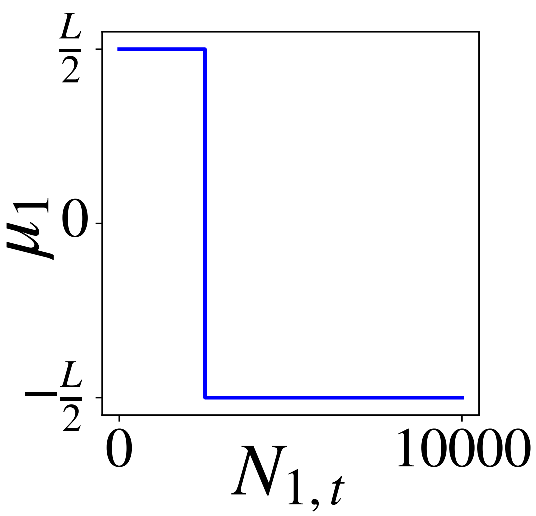
$$R_T(\pi) = \sum_{i \in \mathrm{UP}} \sum_{s=N_{i,T}^{\pi}+1}^{N_{i,T}^{\star}} \mu_i(s) - \sum_{i \in \mathrm{OP}} \sum_{s=N_{i,T}^{\star}+1}^{N_{i,T}^{\pi}} \mu_i(s)$$

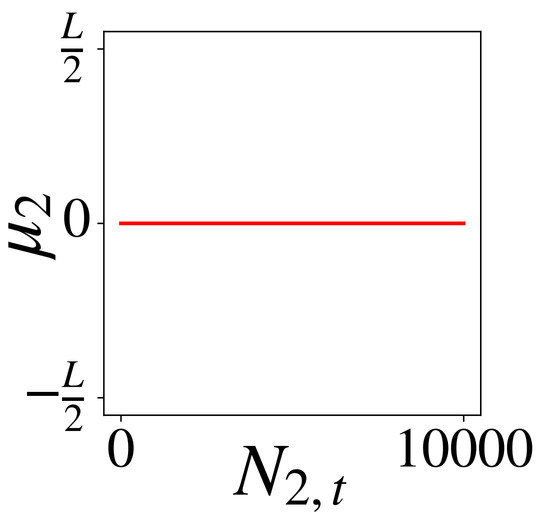
1. **Same (total) number of over-pulls than under-pulls**
2. **The summand in the first sums upper bounded the largest not selected value at the end** $\max_{j \in \mathcal{K}} \mu_j\left(N_{j,T}\right)$
3. **Reward is decreasing:** $\max_{j \in \mathcal{K}} \mu_j\left(N_{j,T}\right) \leq \max_{j \in \mathcal{K}} \mu_j\left(N_{j,t}\right)$
4. **For an over-pulled arm i, the contribution of h_i over-pulls to the regret is bounded by the lemma**

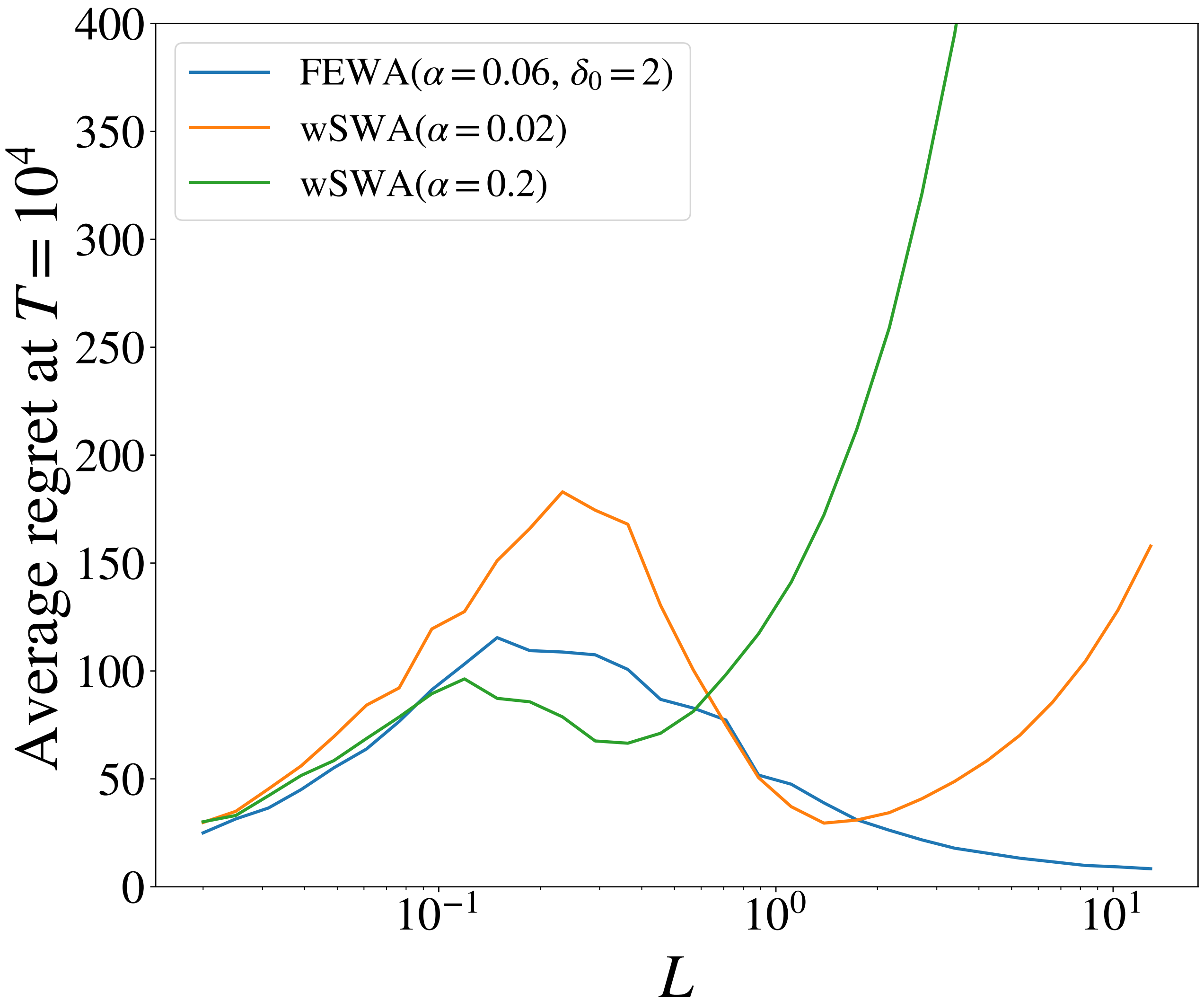
$$h_i\left(\max_{j \in \mathcal{K}} \mu_j\left(N_{j,T}\right) - \overline{\mu}_i^{h_i}\left(N_{i,t}\right)\right) \leq h_i\left(\max_{j \in \mathcal{K}} \mu_j\left(N_{j,t}\right) - \overline{\mu}_i^{h_i}\left(N_{i,t}\right)\right) \leq 4h_i c(h_i, \delta_T) \leq \mathcal{O}\left(\sqrt{h_i \log \frac{1}{\delta_T}}\right)$$

5. **Jensen's inequality** $h_i = T/K$ **(worst case), and get** $R_T(\pi_F) = \mathcal{O}\left(\sqrt{KT \log T}\right)$

Simulations: 2-arms with noise $\sigma = 1$, L (decay)variable







WHERE ARE WE NOW?

Closes the open problem

1 Rotting bandits are not harder than stochastic bandits

- ✅ $\tilde{\mathcal{O}}\left(\sqrt{KT}\right)$ worst-case bound
- ✅ $\tilde{\mathcal{O}}\left(\log T\right)$ problem-dependent bound

2 FEWA, a policy

- ✅ anytime
- ✅ a new data-adaptive window mechanism
- ✅ agnostic/adaptive to L

3 EFF-FEWA, a policy

- ✅ with FEWA's regret guarantees
- ✅ logarithmic space and time complexity

WHAT IS NEXT? FEWA'S LIMITS

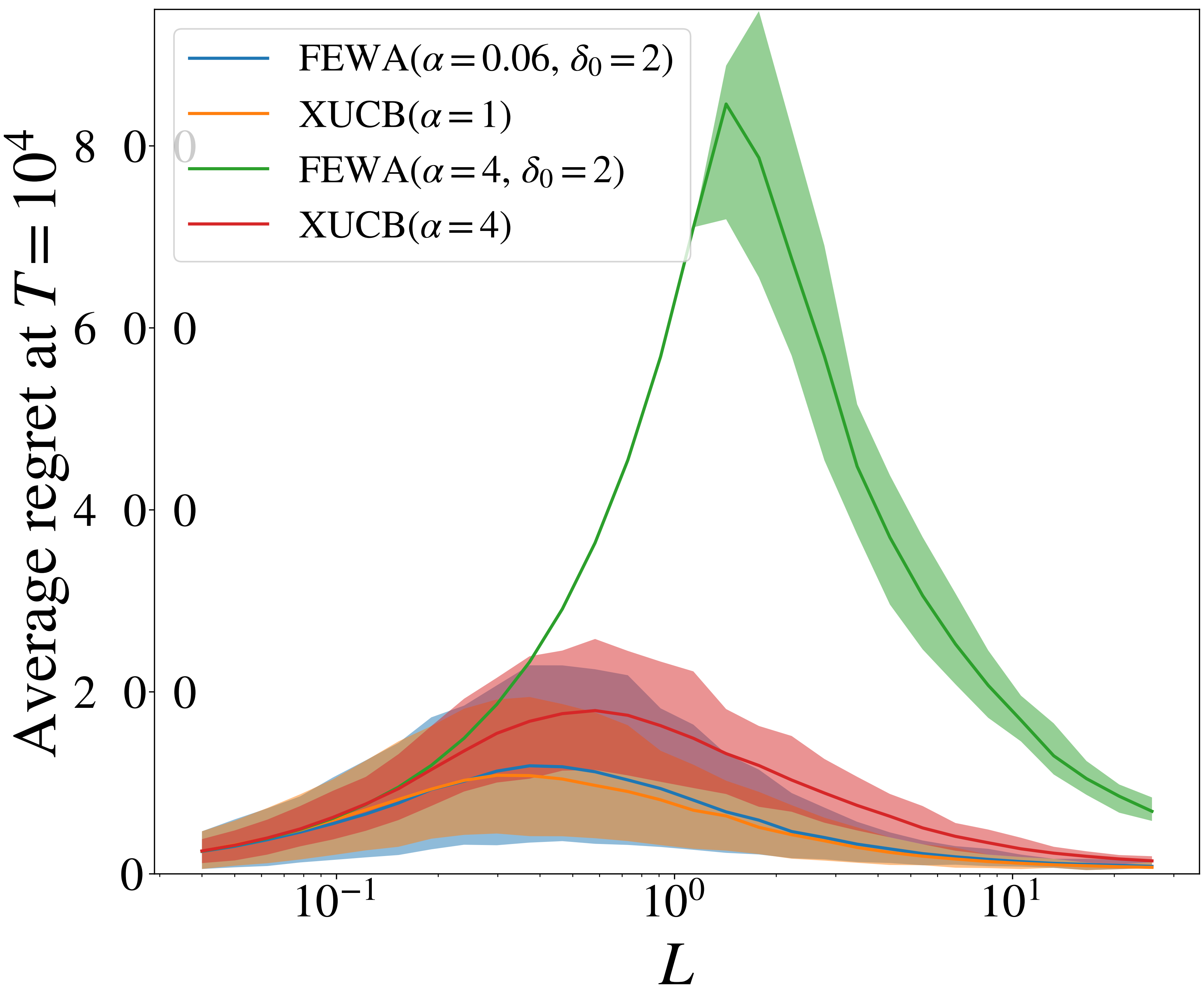
- Complex filtering expanding dynamics
  - Is is possible to have an index policy?
- $\mathcal{O}\left(\log T\right)$ bound $\rightarrow$ 4 times UCB1's bound.
  - Can we do better?
- How about all these different non-stationarity settings?

RAW-UCB: rotting adaptive window UCB

RAW-UCB vs. FEWA

$$\overline{\mu}_i^h(N_{i,t}) \geq \max_{i \in \mathcal{K}} \mu_i(N_{i,t}) - 2c(h, \delta_t)$$

Better than FEWA's "if you pass, you are not too bad"



Stochastic non-stationary bandits

- $\{\mu_i\}$ are functions of round t - restless
- Minimize cumulative regret w.r.t. the optimal strategy: $\sum_{t \leq T} \mu_{i_t^\star}(t) - \mu_{i_t}(t)$
- Unlearnable (w.r.t. a difficult oracle) if μ_i can changes at every round.
- Common settings
  - μ_i is piece-wise stationary with Υ_T pieces *(Garivier & Moulines, 2011)*
  - μ_i has a permitted amount of change V_T *(Besbes et al., 2014)*

$$\sum_{t \leq T} \max_{i \in \mathcal{K}} |\mu_i(t) + 1 - \mu_i(t)| \leq V_T$$

Lower bounds (with the rotting property)

1. Piece-wise stationary problem with $\Upsilon - 1$ equally spaced breakpoints
2. At each breakpoint, the (unknown) best arm is at a distance $\Delta = \mathcal{O}\left(\sqrt{K\Upsilon/T}\right)$ from the others
3. The learner will do at least $\mathcal{O}\left(T/(K\Upsilon)\right)$ mistakes on each suboptimal arm on each batch
4. The learner suffers at least $\mathcal{O}\left(\sqrt{K\Upsilon T}\right)$

Worst value

Υ_T is **fixed**

$$\Upsilon\Delta = \mathcal{O}\left(\sqrt{\frac{K\Upsilon^3}{T}}\right) \leq V_T$$

$$\implies \Upsilon = \mathcal{O}\left(V_T^{2/3}K^{-1/3}T^{1/3}\right)$$

Piece-wise stationary rate:

$$\mathcal{O}\left(\sqrt{K\Upsilon_T T}\right)$$

(Garivier & Moulines, 2011)

Variational budget rate:

$$\mathcal{O}\left(K^{1/3}V_T^{1/3}T^{2/3}\right)$$

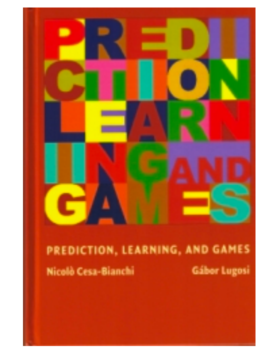
(Besbes et al. , 2014)

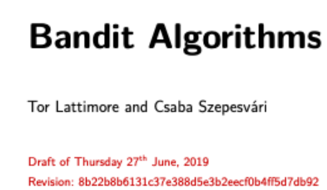
PROBLEM DEPENDENT GUARANTEES

EXP3.S (Auer et al. 2002b), an adversarial algorithm, matches the two minimax rates.
Can we get a problem-dependent bound?

<u>*Theorem 31.2 (Lattimore & Szepesvari, 2019):*</u> Let π a policy suffering $R_T(\mu)$ on a 2-arms stationary bandits problem μ. Then, for T large enough, there exists a piece-wise stationary problem μ' such that π suffers:

$$R_T(\mu') \geq \frac{T}{22 R_T(\mu)}$$




Bandit Algorithms

Tor Lattimore and Csaba Szepesvári

Draft of Thursday 27th June, 2019
Revision: 8b22b8b6131c37e388d5e3b2eecf0b4ff5d7db92

<u>Corollary:</u> No!

Any minimax optimal policy suffers $\mathcal{O}\left(\sqrt{T}\right)$ problem-dependent regret

<u>Why?</u> consider a "quick" increase of the suboptimal arm such that the algorithm cannot notice it.

Quick = Inversely proportional to the sub-optimal arm pulling rate $(R_T(\mu)/\Delta)$.

WHAT ARE THE "BENEFITS" OF BEING STOCHASTIC ?

With no extra properties, what gives? What can be improved?

- Agnostic to Υ_T and V_T (Auer et al., 2019b) - tomorrow!

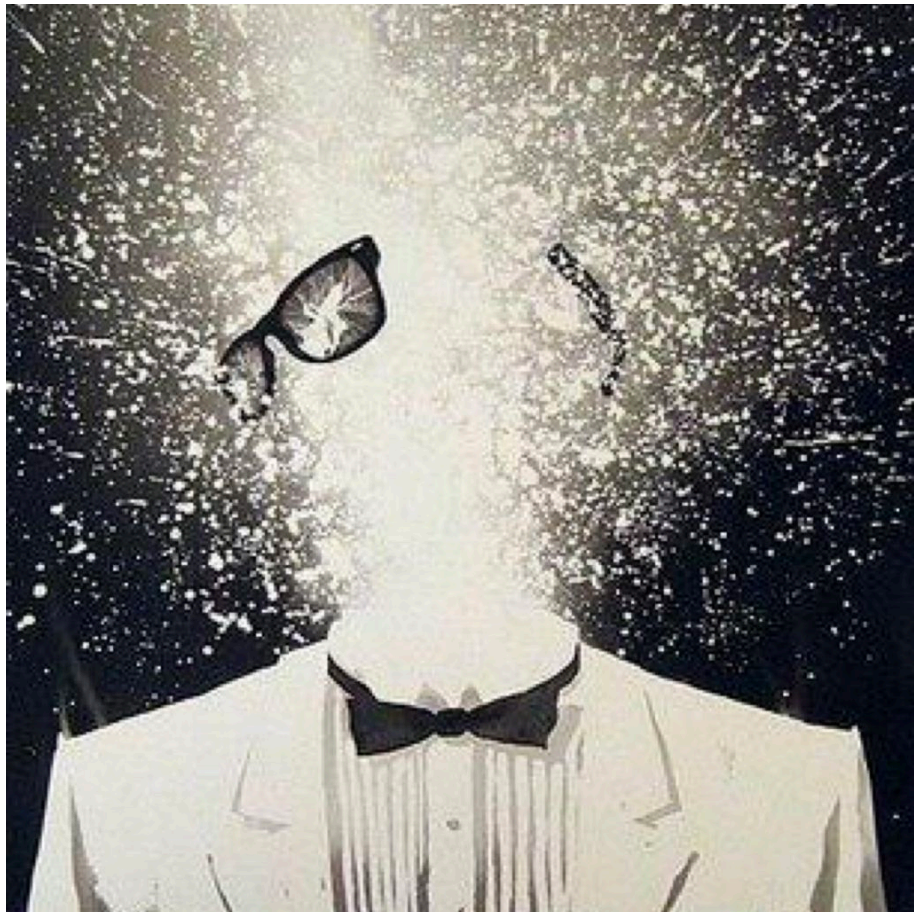
What would we need to recover problem-dependent bound?

Lat/Sze's bible counter-example has 2 properties:

1. The best arm does not change when the suboptimal arm increases
   - …. but for the short time it becomes the best!
   - Note: Mukherjee et Maillard (2019) consider a setup where all the reward moves significantly at each breakpoint. They get PD guarantees!
2. The suboptimal arm is increasing
   - And what in the case when the rewards never increase?

RAW-UCB does not need to know in which setup it is

RAW-UCB without knowing T, V_T nor Υ_T



Variational budget

$$\mathbb{E}\left[R_T\left(\pi_{\mathrm{R}}\right)\right] \leq \tilde{\mathcal{O}}\left(\sigma^{2/3}K^{1/3}V_T^{1/3}T^{2/3}\right)$$

Piecewise stationary

$$\mathbb{E}\left[R_T\left(\pi_{\mathrm{R}}\right)\right] \leq \tilde{\mathcal{O}}\left(\sigma\sqrt{K\Upsilon_T T}\right)$$

$$\mathbb{E}\left[R_T\left(\pi_{\mathrm{R}}\right)\right] \leq \sum_{i\in\mathcal{K}}\sum_{k=1}^{\Upsilon_T}\frac{32\sigma^2\log T}{\Delta_{i,h}} + \mathcal{O}\left(\sqrt{\log T}\right)$$

problem-dependent bound!

SKETCH OF PROOFS

Lemma 3. *On favorable event ξ_t, if* RAW-UCB *selects an arm $i \in \mathcal{K}$ at round t, for any $h \leq N_{i,t}$, the average of its h last pulls cannot deviate significantly from the best available arm at that round, i.e.,*

Which one is useful? $$\overline{\mu}_i^h(\pi_R, t) \geq \max_{i \in \mathcal{K}} \mu_i(t) - 2c(h, \delta_t).$$

Piecewise stationary bandits

- *We choose h such that we include all the sample from the current stationary batch*
- *On each batch, the proof is then similar to UCB1's.*

SKETCH OF PROOFS

Lemma 3. *On favorable event ξ_t, if* RAW-UCB *selects an arm $i \in \mathcal{K}$ at round t, for any $h \leq N_{i,t}$, the average of its h last pulls cannot deviate significantly from the best available arm at that round, i.e.,*

Which one is useful? $\overline{\mu}_i^h(\pi_R, t) \geq \max_{i \in \mathcal{K}} \mu_i(t) - 2c(h, \delta_t).$

Variational budget bandits

1. **We design Υ batches of equal length**
2. **We choose h such that we include all the sample from the current designed batch.**
3. **We split the regret into two sums**
   The regret due to the variance of $\overline{\mu}_i^h(\pi_X, t)$ (Lemma 3) : $\tilde{\mathcal{O}}(\sqrt{KT\Upsilon})$
   The regret due to the bias of $\overline{\mu}_i^h(\pi_X, t)$ compared to the current value: $\tilde{\mathcal{O}}(V_T T/\Upsilon)$
4. **We choose $\Upsilon = \tilde{\mathcal{O}}(V_T^{2/3} T^{1/3} K^{-1/3})$ adequately**

RAW CONCLUSIONS

1 **Rotting property makes restless bandits easier**

✅ $\mathcal{O}(\log T)$ ... bound

2 **UCB-lik... ...ting restless bandits**

✅ No ... exploration, no pass... change-detection routine
✅ Ga... ...ent and minimax bounds
✅ Ag... V_T, T

3 **RAW-U... ...d OR restless ... bandits**

✅ with the sa...
❌ rested AND re... ...ble