SEQUENTIAL MACHINE LEARNING FOR ADAPTIVE EDUCATIONAL SYSTEMS

PhD defense of Julien SEZNEC, the 15th of December 2020

Supervised by:
Michal VALKO (Inria Lille / Deepmind)
Alessandro LAZARIC (Inria Lille / FAIR)
Jonathan BANON (Lelivrescolaire.fr)

inventors for the digital world

lelivrescolaire.fr

OUTLINE

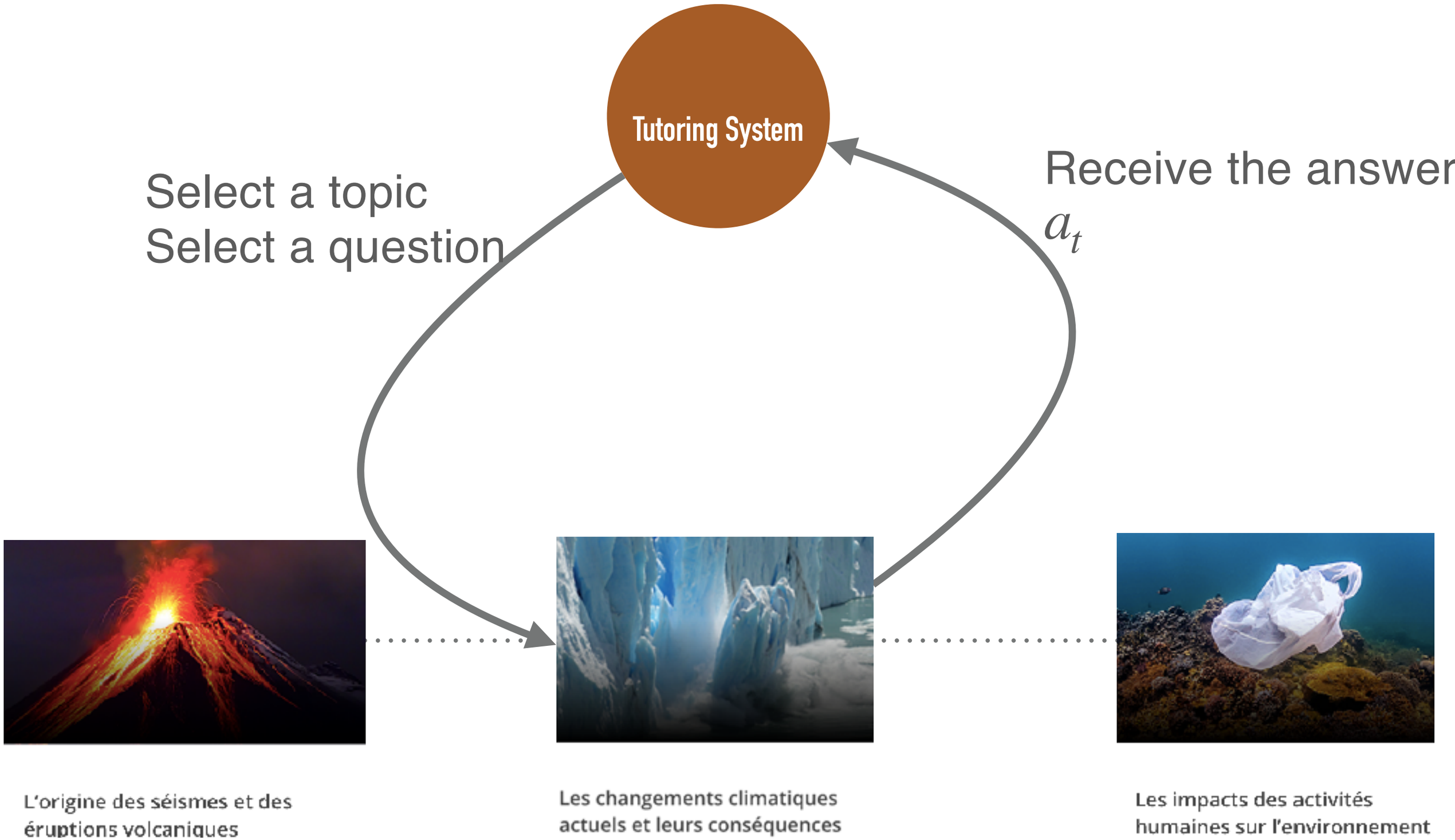
1. Bandits for Adaptive Educational Systems

2. Rested Rotting Bandits

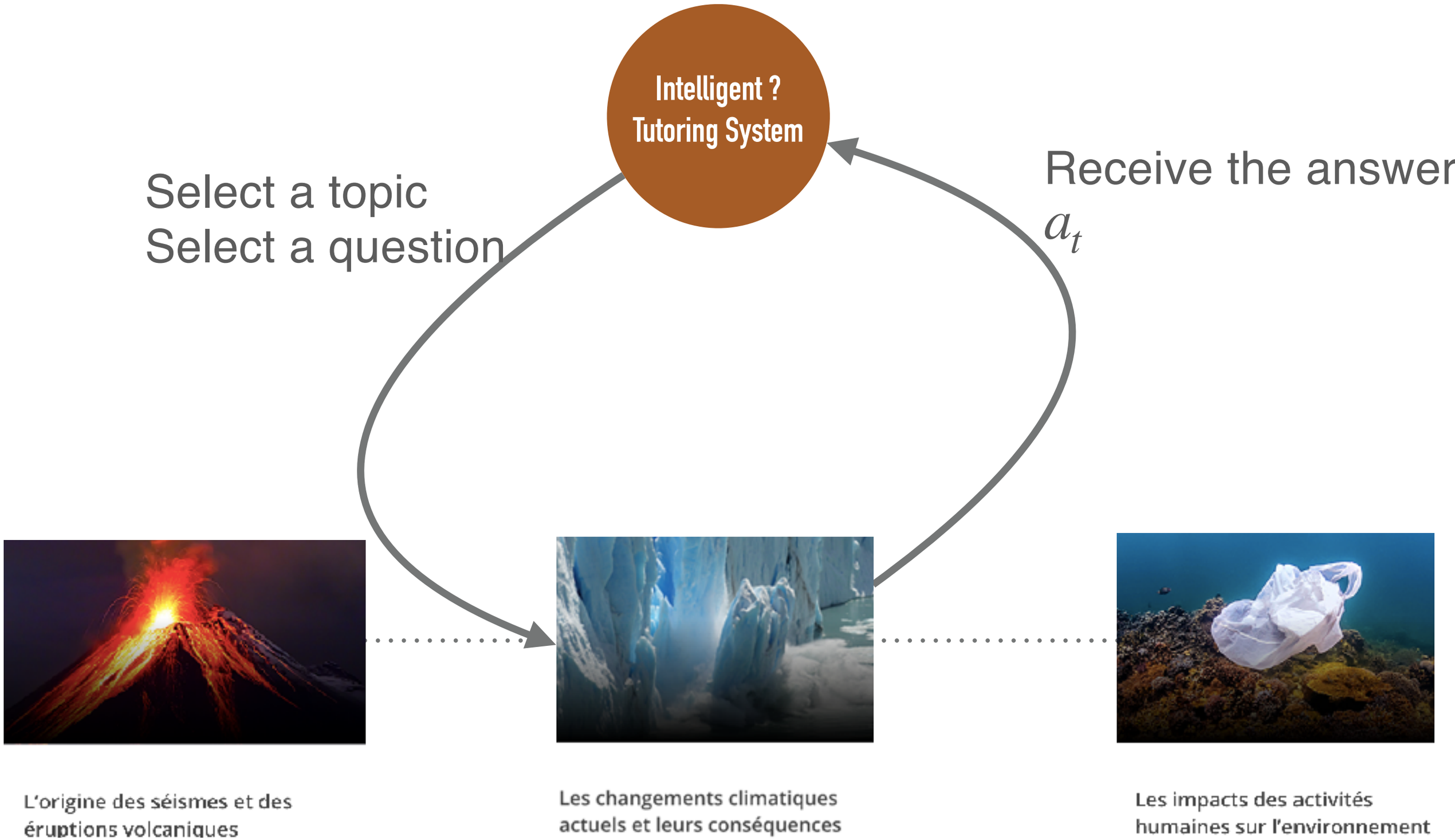
3. Restless Rotting Bandits

AFTERCLASSE

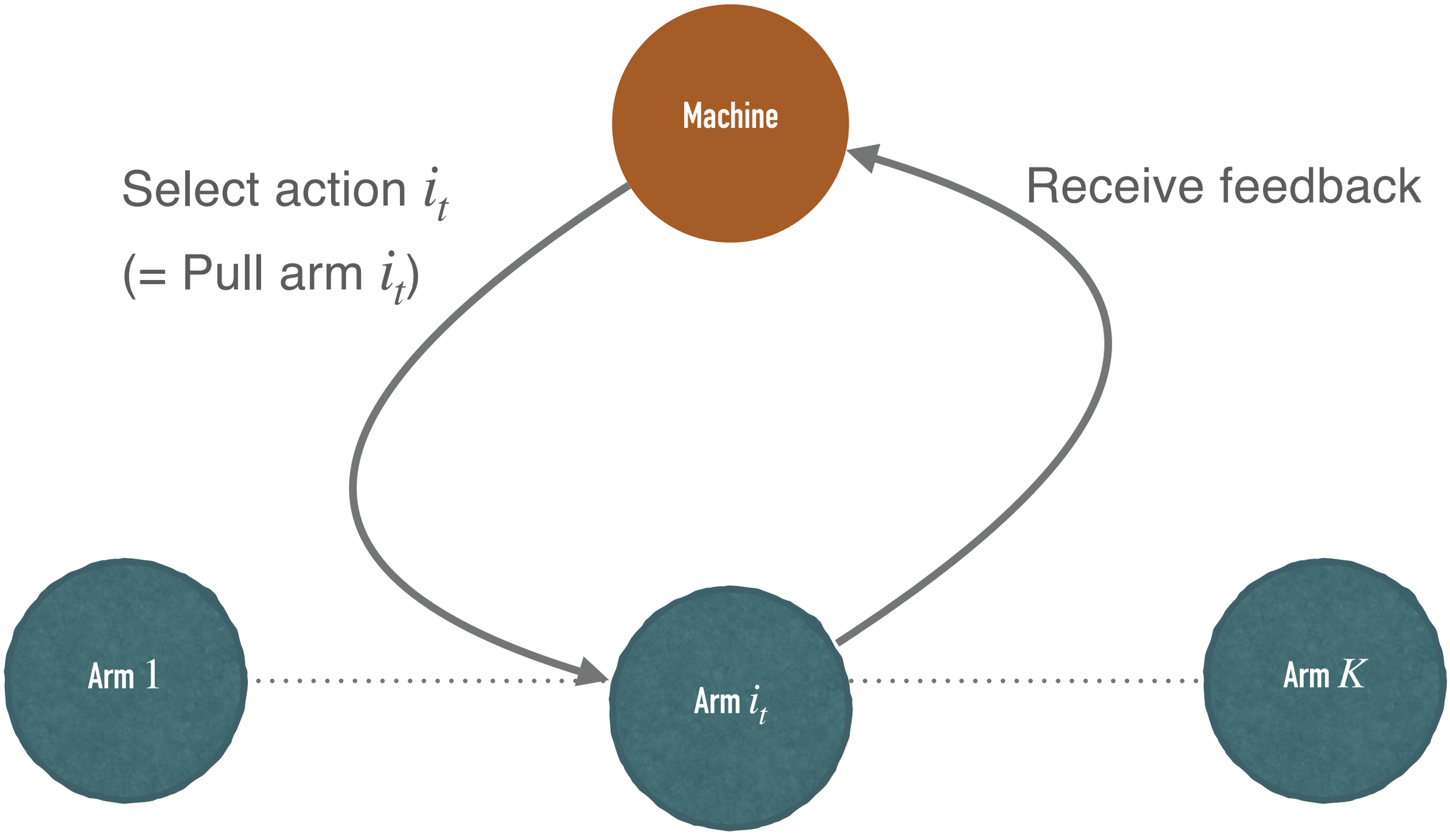
Sequential Adaptive Educational Systems



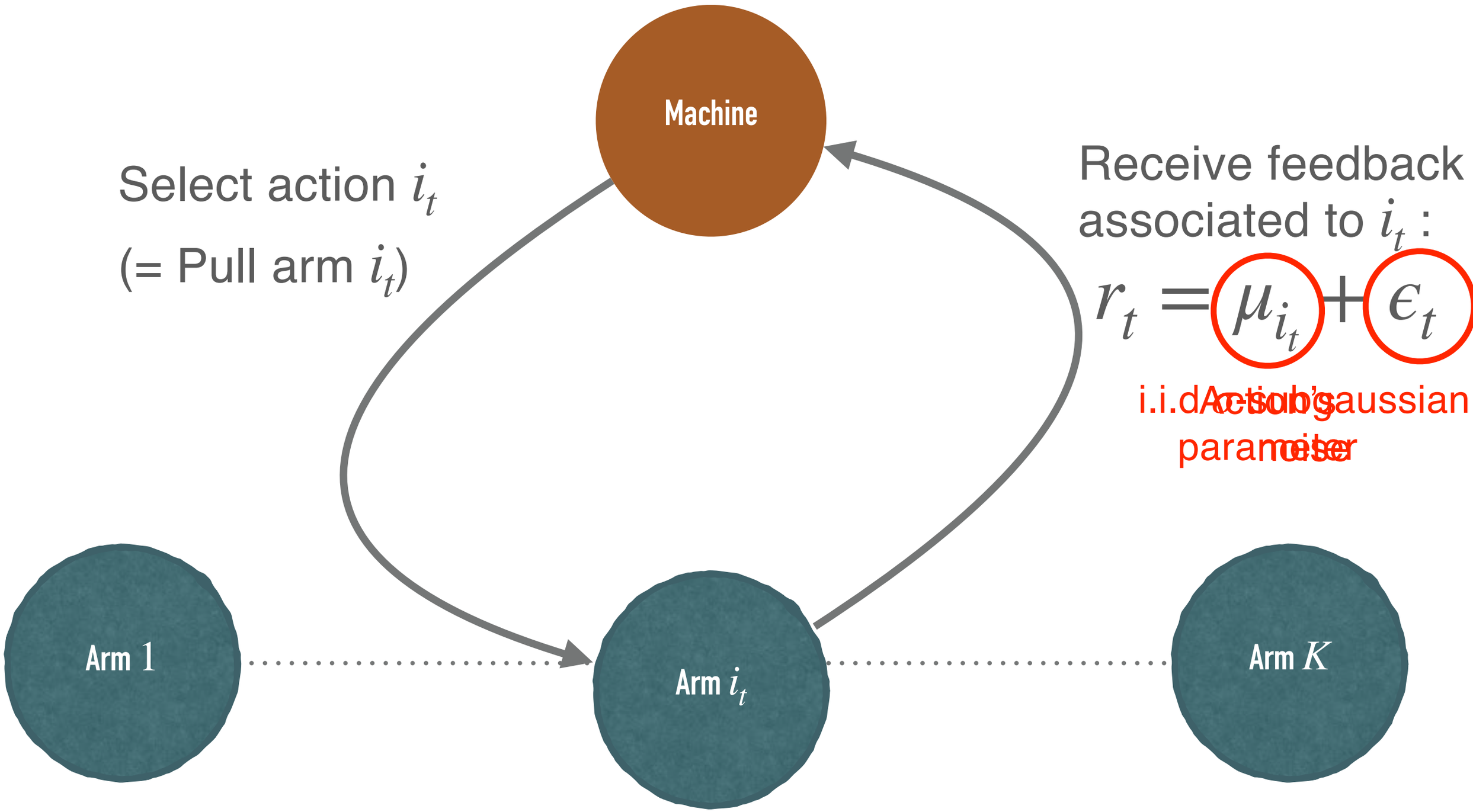
Sequential Adaptive Educational Systems



Sequential Exploration: Bandits Feedback



Stationary Bandits: Feedback



Stationary Bandits: Objective

- Standard view : Feedback = reward
- Objective : Maximize cumulative reward in expectation at the horizon T

$$J_T(\pi) = \mathbb{E}\left[\sum_{t<T} r_t\right] = \mathbb{E}\left[\sum_{t<T} \mu_{i_t}\right]$$

- Optimal oracle policy

$$\pi^\star(t) \in \operatorname{argmax}_{i \leq K} \mu_i$$

- Regret

$$R_T(\pi) = J_T(\pi^\star) - J_T(\pi) = \sum_{i \in \mathcal{K}} N_{i,T} \Delta_i$$

$\Delta_i = \max_j \mu_j - \mu_i$

$N_{i,T}$ Number of pulls of arm i at round T

Upper Confidence Bound Policy

Select the arm with the largest index:

$$\hat{\mu}_i + \sqrt{\frac{2\sigma^2 \log 1/\delta}{N_{i,t}}}$$

Expected average reward + cost of information

Empirical average + bonus for confidence level

$$\delta = \max\left(\frac{KN_{i,t}}{t}, 1\right)$$

Optimal problem-independent rate:

$$\Theta\left(\sigma\sqrt{KT}\right)$$

Optimal problem-dependent rate:

$$\sum_{i,\Delta_i>0} \frac{2\sigma \log(T)}{\Delta_i}$$

Bandits in Educational Systems

1. What is the objective ? What is the reward ?

2. What is the « best » action ?

3. How do actions impact observations ?

4. What can we learn with ~100 samples ?

Bandits In Educational Systems

Clément et al. (2016)	EXP3	Target the topic with the largest improvement	$r_t = \frac{1}{h}\sum_{i=0}^{h-1}\left(a_{t-i} - a_{t-h-i}\right)$
Melesko et al. (2019)	UCB	Target the most difficult topic	$r_t = -a_t$
Lin et al. (2020)	Thomson sampling	Target the most difficult topic	$r_t = -a_t$

Bandits In Educational Systems

Clément et al. (2016)	EXP3	Target the topic with the largest improvement	$r_t = \frac{1}{h}\sum_{i=0}^{h-1}\left(a_{t-i} - a_{t-h-i}\right)$
Melesko et al. (2019)	UCB	Target the most difficult topic	$r_t = -a_t$
Lin et al. (2020)	Thomson sampling	Target the most difficult topic	$r_t = -a_t$

OUTLINE

1. Bandits for Adaptive Educational Systems

2. Rested Rotting Bandits

3. Restless Rotting Bandits

When Bandits Go Rotting …

$$a_t \uparrow \Longrightarrow \quad r_t = -\, a_t \downarrow$$

L'origine des séismes et des éruptions volcaniques

Les changements climatiques actuels et leurs conséquences

Les impacts des activités humaines sur l'environnement

Rested vs. Restless Bandits

Rested :

- Actions trigger non-stationarity
- No action = no change = rested
- $\mu_i(N_{i,t})$

Restless :

- Changes happen independently of the actions
- Action can change when they are not selected
- $\mu_i(t)$

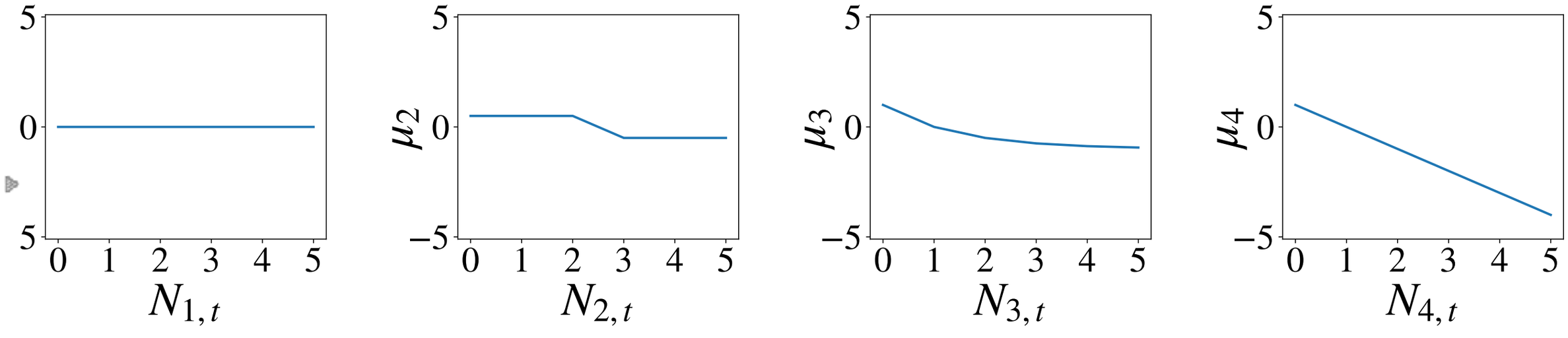
Rested Rotting Bandits are …

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: $J_T(\pi) = \mathbb{E}\left[\sum_{t \leq T} r_t\right]$

… with rotting arms

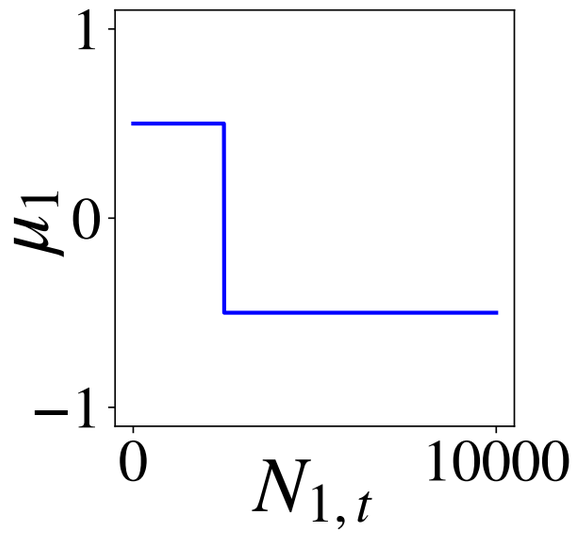
- $\{\mu_i\}$ are ***non-increasing*** functions of $N_{i,t}$ the ***number of pulls of arm*** i at time t
- $L \triangleq \max_{i \in K} \max_{n \leq T} \mu_i(n) - \mu_i(n+1)$

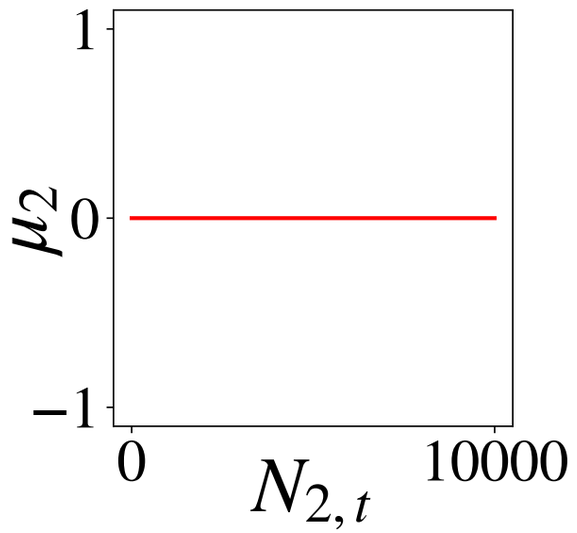


Oracle & Regret

Optimal Oracle:

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

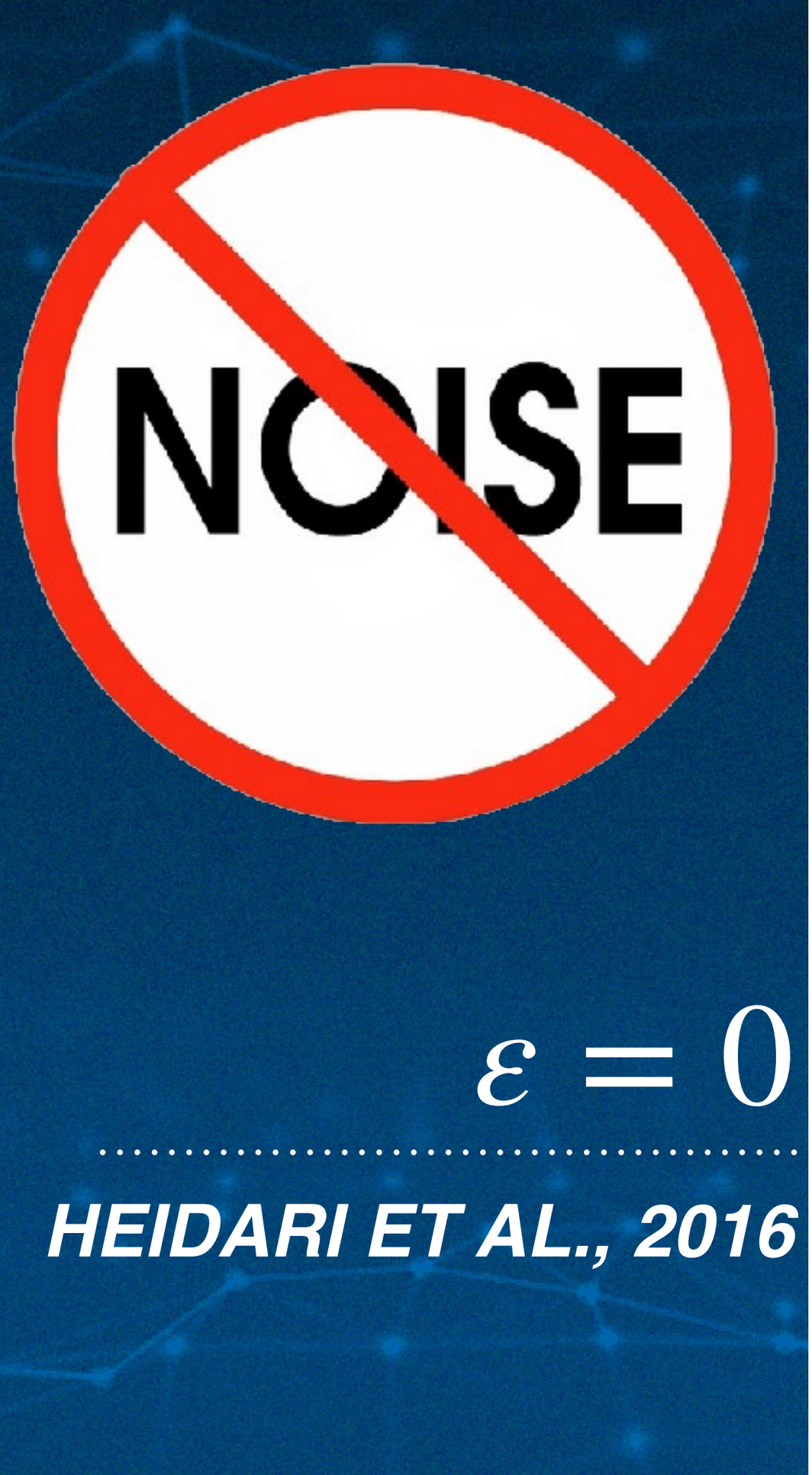




$$R_T(\pi) = J_T(\pi^\star) - J_T(\pi)$$

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

$$\begin{cases} \mathrm{UP} : \text{underpulled arms by } \pi \text{ w.r.t } \pi^\star \\ \mathrm{OP} : \text{overpulled arms by } \pi \text{ w.r.t } \pi^\star \end{cases}$$

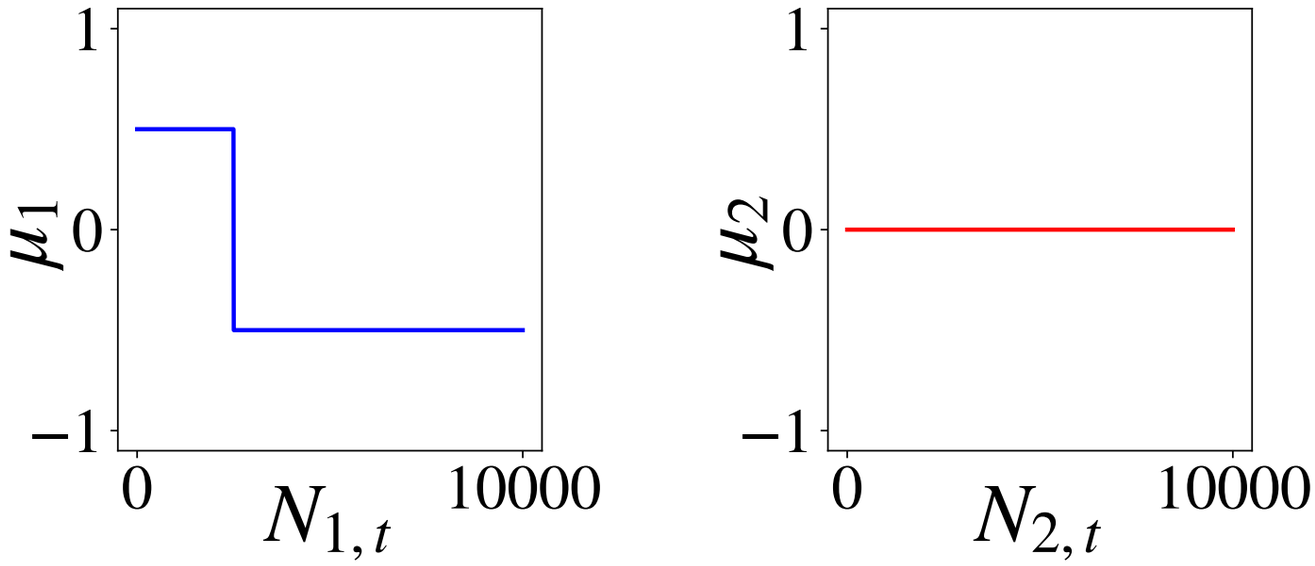


Optimal Oracle:
$\pi_O(t) = \arg\max_i \mu_i \left(N_{i,t}\right)$

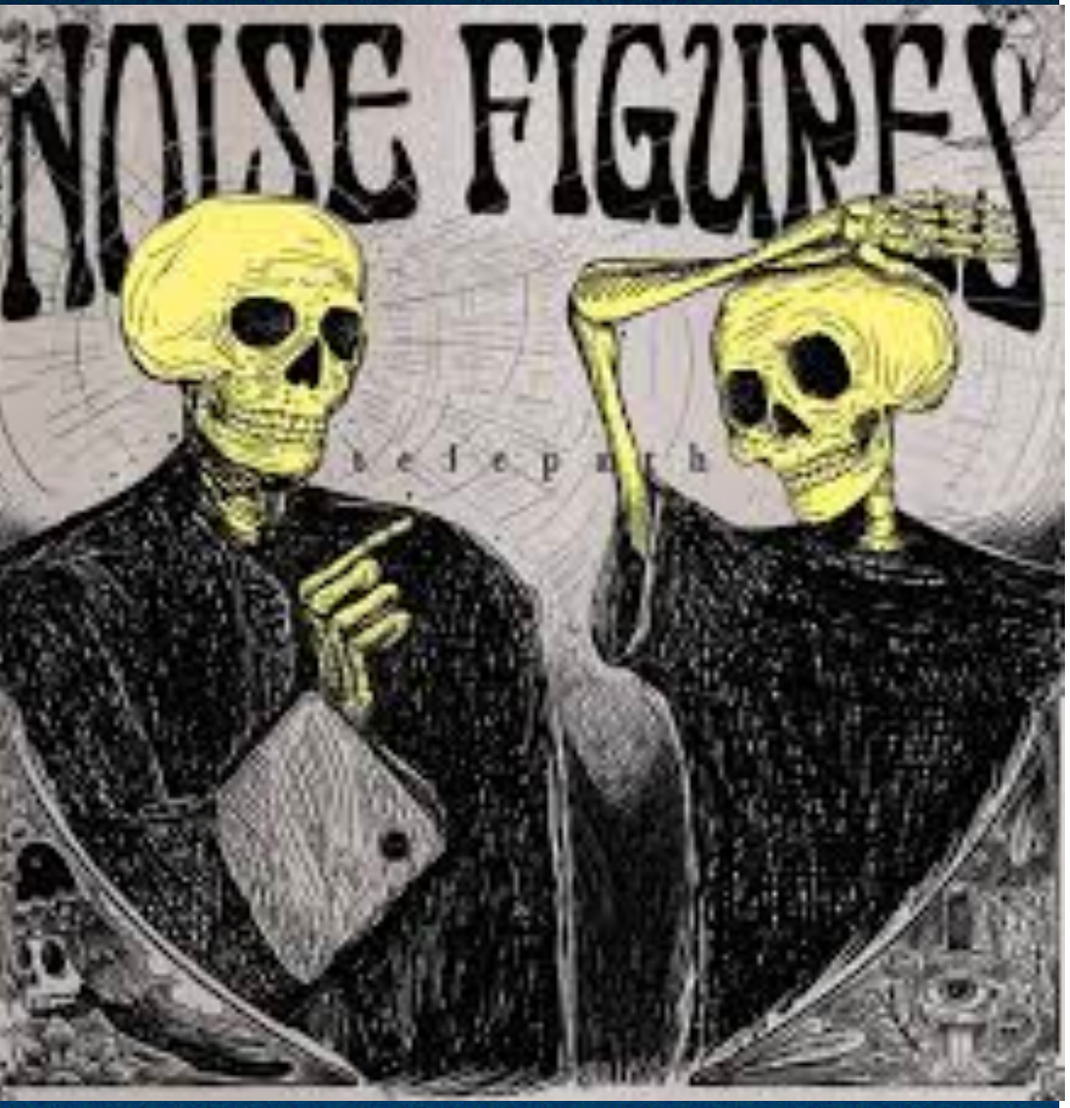
Online bandit policy:
$\pi_B(t) = \arg\max_i \mu_i \left(N_{i,t} - 1\right)$

Minimax regret rate: $R_T(\pi_B) \leq (K-1)\,L$

$\rightarrow$ At most,1 mistake per arm.



Rested : New mistake cancels the old one !



NOISE

LEVINE ET AL., 2017

NOISE = AVERAGE

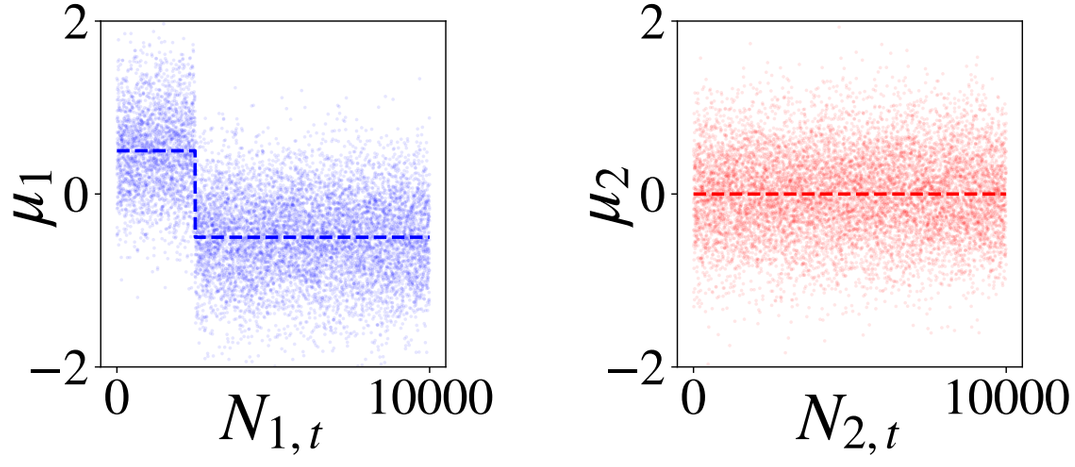
(Which samples? How many?)

wSWA [LEVINE, 2017]

Algorithm 3 SWA (Levine et al., 2017)

Input: K, L, T, σ

1: $h \leftarrow \widetilde{\mathcal{O}}\left(\left(\frac{\sigma T}{KL}\right)^{2/3}\right)$
2: **for** $t \leftarrow Kh+1, Kh+2, \ldots$ **do**
3: SELECT : $\arg\max_{i \in \mathcal{K}} \widehat{\mu}_i^h(N_{i,t})$
4: **end for**



Average of the h last samples

Regret due to bias: $\widetilde{O}\left(LKh\right)$

Regret due to variance : $\widetilde{O}\left(\sigma T\sqrt{h}^{-1}\right)$

Worst case regret : $\widetilde{O}\left(K^{1/3}T^{2/3}\right)$

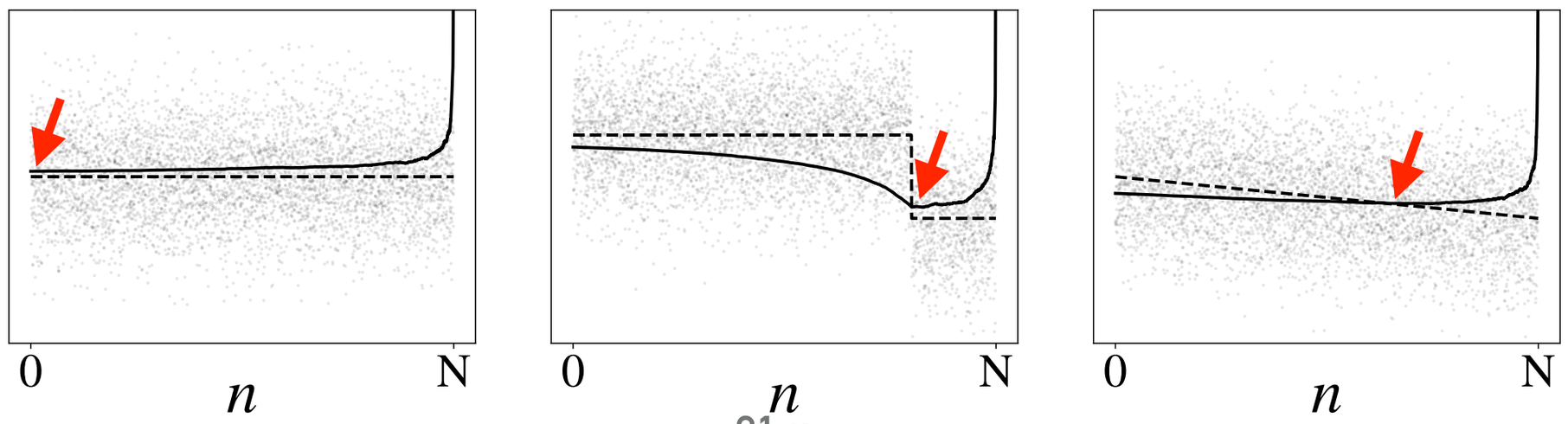
ROTTING ADAPTIVE WINDOW UCB

Algorithm 5 RAW-UCB

Input: K, σ, α

1: **for** $t \leftarrow K+1, K+2, \ldots$ **do**
2: $\quad \delta_t \leftarrow \frac{1}{t^\alpha}$
3: $\quad$ SELECT : $\arg\max_{i \in K} \min_{h \leq N_{i,t}} \left(\widehat{\mu}_i^h(N_{i,t}) + c(h, \delta_t) \right)$
4: **end for**

- Rewards are decreasing: $\widehat{\mu}_i^h\left(N_{i,t}\right) + c(h, \delta_t)$ is a UCB for the future value
- RAW-UCB selects the minimum (tightest) one (on h) as index of the arm



Upper Bounds

Worst-case upper bound

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

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

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

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 RAW-UCB

Upper Bounds

Worst-case upper bound

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

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

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

Comparison w/ wSWA

Pure worst-case strategy

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

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

Proof Sketch

1. Decompose the regret

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

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

μ_T^+ : largest underpulled values

$\sim$ "number of mistakes" $\times$ "average gap"

2. Use the adaptive window mechanism to get

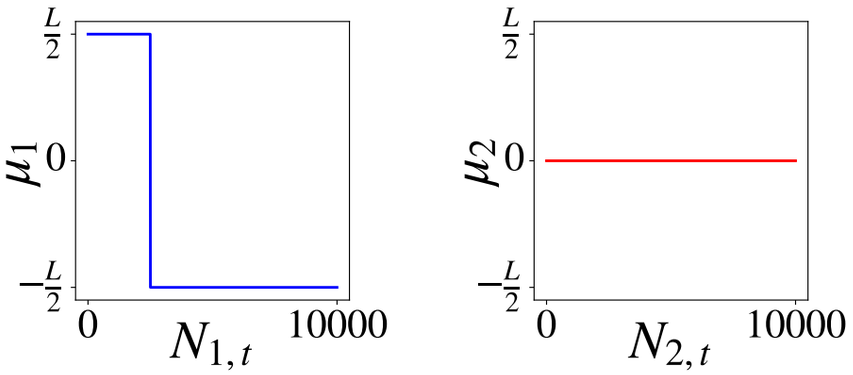
$$\text{"number of mistakes"} \sim \frac{\log T}{\text{"average gap"}^2}$$

3. Work on the gap to be problem-dependent

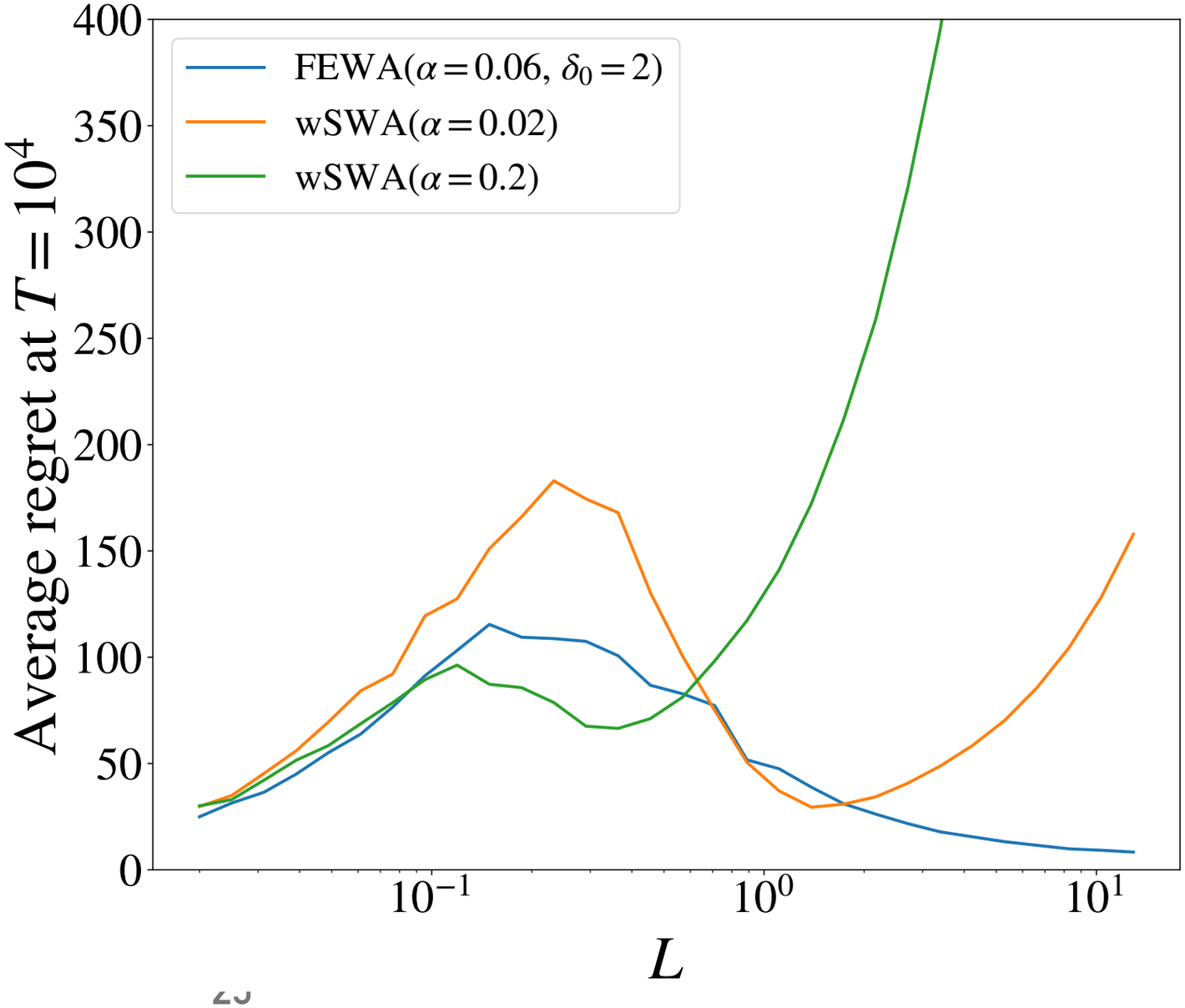
"number of mistakes" $\rightarrow$ "maximal number of mistakes with high probability"

$$\mu_T^+ \rightarrow \min_i \min_{s \leq N^{\star}_{i,T}} \mu_i(s)$$

Experimental Results



$$\sigma = 1 \; ; \; L$$



RESTED CONTRIBUTIONS

Algorithm	Sliding Window Average (Levine et al., 2017)	FEWA & RAW-UCB
Average(s)	Fixed single window	Multiple windows
Problem Independent Bound	$\tilde{\mathcal{O}}\left(L^{1/3}\sigma^{2/3}K^{1/3}T^{2/3}\right)$	$\tilde{\mathcal{O}}\left(\sigma\sqrt{KT}\right)+\mathcal{O}\left(KL\right)$
Problem-Dependent Bound	✗	$\mathcal{O}\left(\sum_{i\in\mathcal{K}}\frac{\log T}{\Delta_i^h}\right)$
Knowledge	T L σ	σ
Complexity per round	$\mathcal{O}\left(K^{1/3}T^{2/3}\right)$	$\mathcal{O}\left(T\right)$ efficient algo $\mathcal{O}\left(K\log T\right)$

OUTLINE

1. Bandits for Adaptive Educational Systems

2. Rested Rotting Bandits

3. Restless Rotting Bandits

Non-Stationary Restless Bandits

- $\{\mu_i\}$ are functions of round t - restless
- Minimize cumulative regret w.r.t. the optimal strategy: $\sum_{t \le T} \mu_{i_t^\star}(t) - \mu_{i_t}(t)$
- Unlearnable if μ_i can change significantly 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)*

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

Piece-wise stationary rate:

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

(Garivier & Moulines, 2011)

Variational budget rate:

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

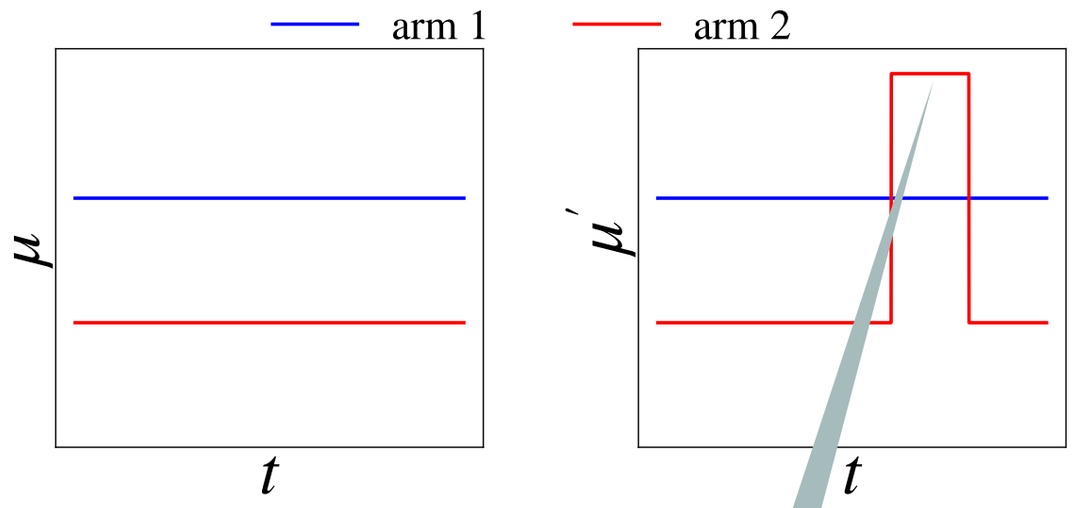
(Besbes et al. , 2014)

Problem Dependent Guarantees

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

Theorem (Garivier et al. 2011): **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:**

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



Quick ***increase***

$$\tau = \frac{T}{R_T(\mu)/\Delta}$$

Corollary : No !

$\mathcal{O}\left(\sqrt{T}\right)$ **anywhere ⇨** $\mathcal{O}\left(\sqrt{T}\right)$ **everywhere**

RESTLESS ROTTING

Do we need to spend $\mathcal{O}\left(\sqrt{T}\right)$ exploration budget when rewards *decay* ?

RAW-UCB on Rotting Restless Bandits

RAW-UCB without knowing T, V_T nor Υ_T

Rotting Variational budget

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

minimax rate

Rotting Piecewise stationary

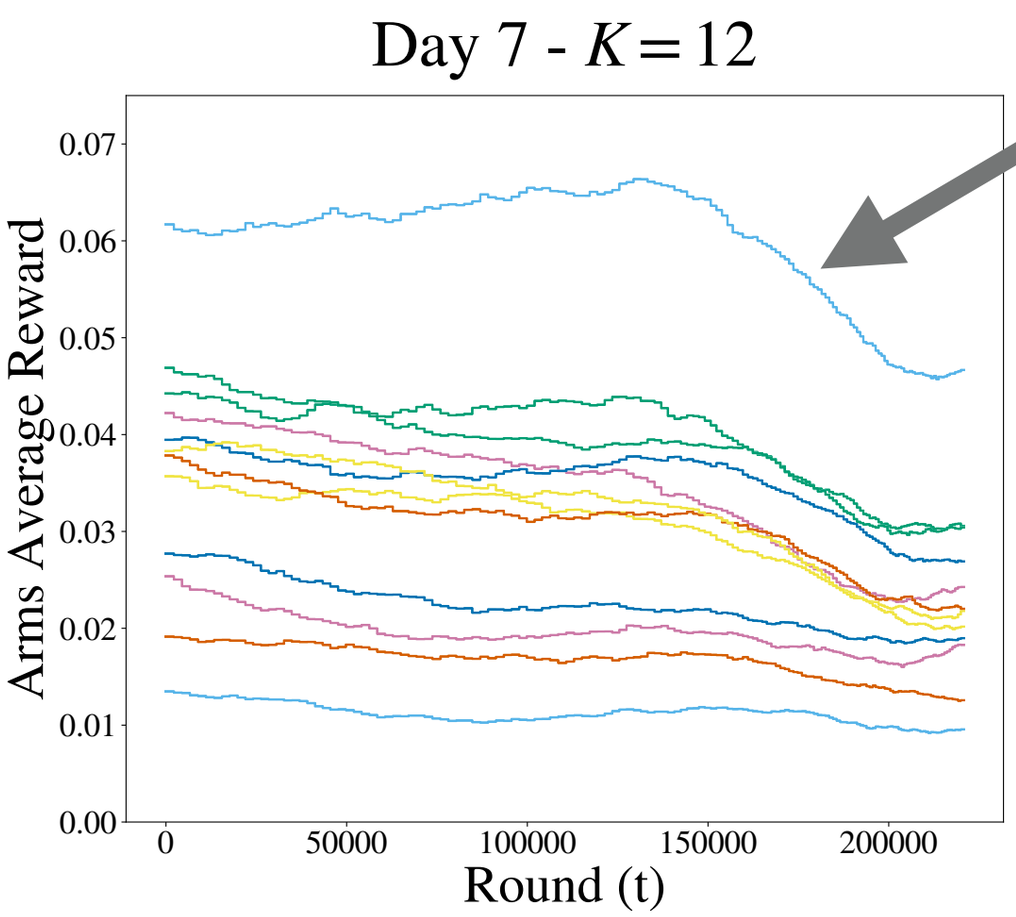
$$\mathbb{E}\left[R_T\left(\pi_{\mathrm{R}}\right)\right] \leq \tilde{\mathcal{O}}\left(\sigma\sqrt{K\Upsilon_T T}\right)$$

minimax rate

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

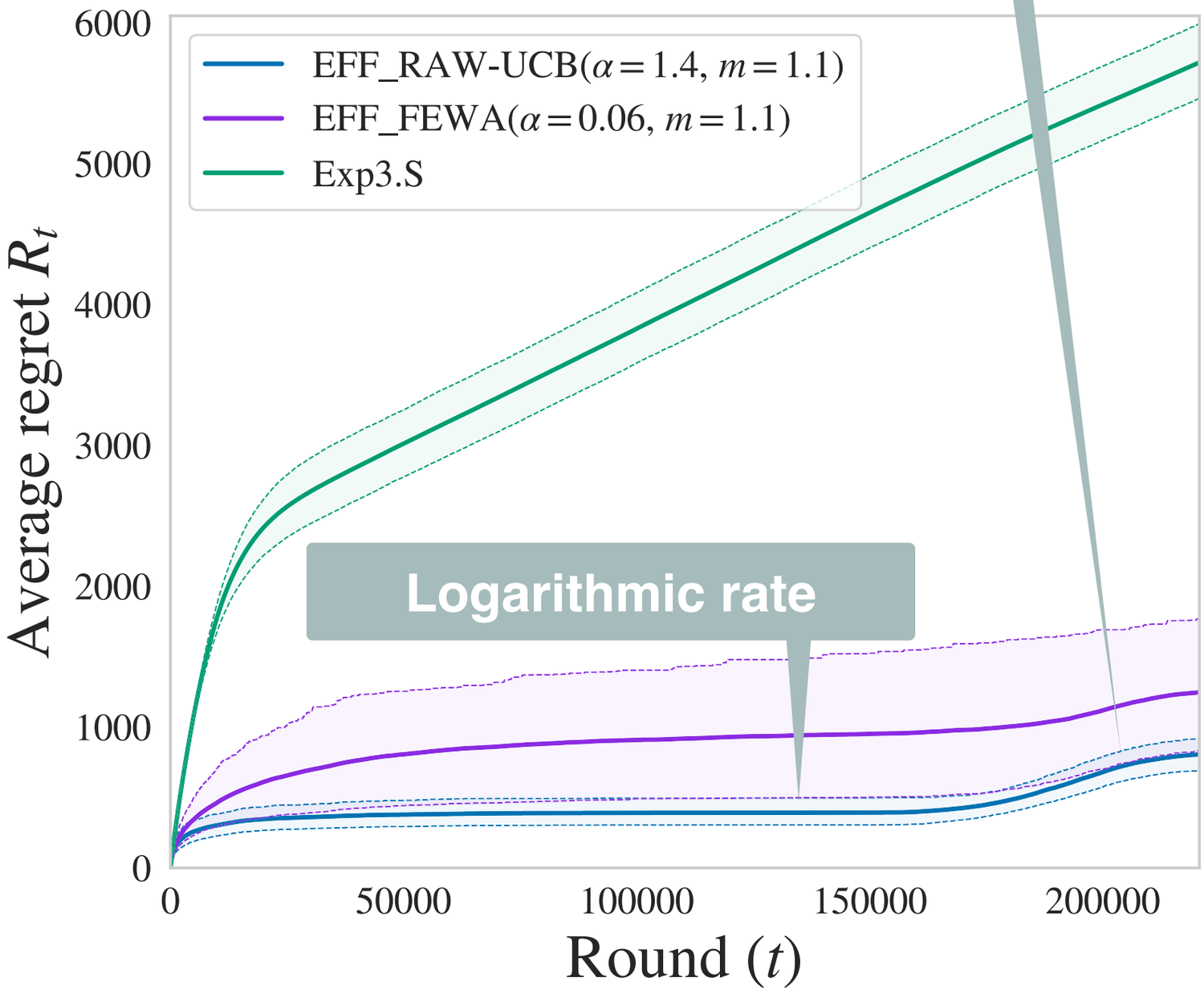
problem-dependent bound!

YAHOO! BENCHMARK



Decaying trend in the probability of click on an article from noon to 6am.

Significant decay : exploration



Restless Contributions

1 Rotting property makes restless bandits easier

✅ $\mathcal{O}\left(\log T\right)$ problem-dependent bound

2 UCB-like index is sufficient for rotting bandits

✅ No random exploration, no passive forgetting, no change-detection routine

✅ Near-optimal gap-dependent and minimax bounds

✅ Agnostic to Υ_T, V_T, T

3 RAW-UCB solves rested OR restless rotting bandits

✅ with the same tuning

❌ rested AND restless rotting bandits are **incompatible**

CONTRIBUTIONS TO EDUCATIONAL SYSTEMS

1. What is the objective ?	Target the most difficult topic	
2. What is the « best » action ?	✅ Optimal policy selects different actions when the student progresses	
3. How does action impact observation?	✅ Model the learning of the student	
4. Can we learn something relevant with ~100 samples?	✅ Not harder than stationary bandits	

CONTRIBUTIONS TO EDUCATIONAL SYSTEMS

1. What is the objective ?	Target the most difficult topic	❌ Target the easiest topic until it is mastered
2. What is the « best » action ?	✅ Optimal policy selects different actions when the student progresses	
3. How does action impact observation?	✅ Model the learning of the student	
4. Can we learn something relevant with ~100 samples?	✅ Not harder than stationary bandits	❌ RAW-UCB = Round Robin

THANK YOU !

Rotting Bandits are not Harder than Stochastic ones,
Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, Michal Valko,
Artificial Intelligence and Statistics (2019), (with oral presentation)

A Single Algorithm for both Rested and Restless Rotting Bandits,
Julien Seznec, Pierre Ménard, Alessandro Lazaric, Michal Valko,
Artificial Intelligence and Statistics (2020)