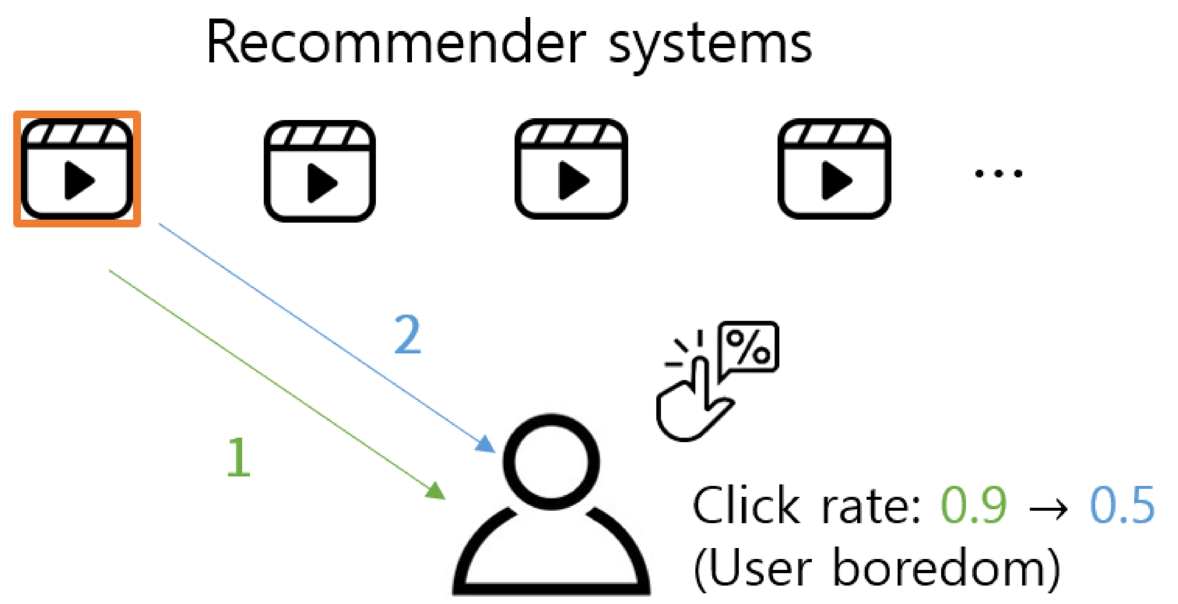
An Adaptive Approach for Infinitely Many-armed Bandits under Generalized Rotting Constraints

Jung-hun Kim[1], Milan Vojnović[2], Se-Young Yun[3]
[1]*Seoul National University*, [2]*London School of Economics*, [3]*KAIST*

Introduction

We consider a fundamental sequential learning problem in which an agent must play one arm at a time *from an infinite set of arms with rotting rewards*, where the mean reward of a selected arm may decrease at each play of the arm.

Applications.



- Content recommendation systems where click rates of items may decrease because of user boredom when watching the same content.
- Clinical trials in which the efficacy of a medicine may decrease due to drug tolerance when a patient takes the same medicine several times.

Problem Statement

- There are infinitely many arms in $\mathcal{A}$.
- The stochastic reward gained by pulling arm a_t at time t is defined as

$$r_t = \mu_t(a_t) + \eta_t.$$

- The initial mean rewards $\{\mu_1(a) \in [0,1]\}_{a\in\mathcal{A}}$ are i.i.d. rv following

$$\mathbb{P}(\mu_1(a) > 1-x) = \mathbb{P}(\Delta_1(a) < x) = \Theta(x^\beta).$$

- At each time t, the mean rewards of pulled arm a_t is updated as

$$\mu_{t+1}(a_t) = \mu_t(a_t) - \rho_t.$$

The values of rotting rates of pulled arms, $\{\rho_t\}_{t\in[T-1]}$, are assumed to be determined by an adversary. We consider two cases for rotting rates:

- *Slow rotting case*: for given $V_T \geq 0$, $\sum_{t=1}^{T-1}\rho_t \leq V_T$.
- *Abrupt rotting case*: for given $S_T \in [T]$, $1 + \sum_{t=1}^{T-1}\mathbb{1}(\rho_t \neq 0) \leq S_T$.

The objective is to find a policy that minimizes the following expected cumulative regret

$$\mathbb{E}[R^\pi(T)] = \mathbb{E}\left[\sum_{t=1}^{T}(1-\mu_t(a_t))\right].$$

Algorithm

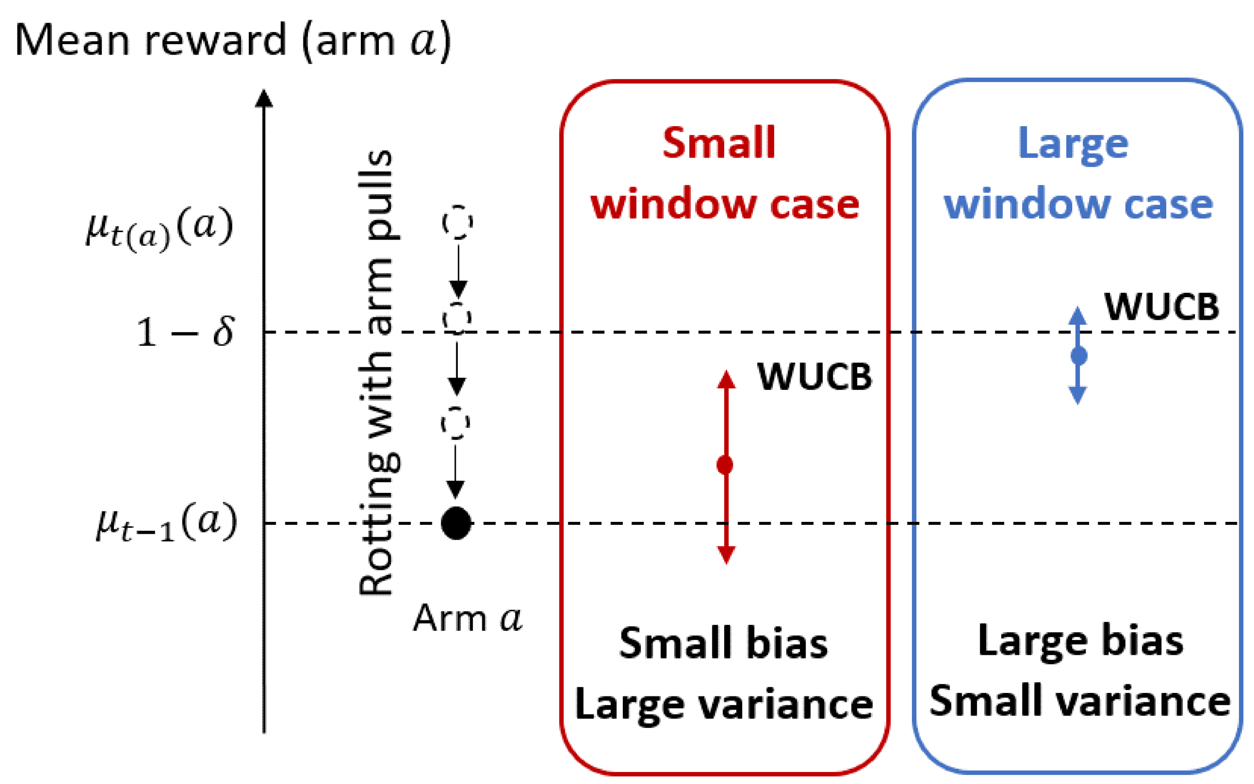
We propose an algorithm, referred to as UCB-Threshold with Adaptive Sliding Window. Here we explain how the algorithm works.

- The algorithm first selects an arbitrary new arm $a \in \mathcal{A}$.
- Then for window-UCB index of the algorithm, we define

$$WUCB(a,t_1,t_2,T) = \widehat{\mu}_{[t_1,t_2]}(a) + \sqrt{12\log(T)/n_{[t_1,t_2]}(a)}.$$

- Then the algorithm pulls the arm consecutively until the following threshold condition is satisfied:

$$\min_{s\in\mathcal{T}_t(a)} WUCB(a,s,t-1,T) < 1-\delta.$$



- For slow rotting (V_T), we set $\delta = \delta_V(\beta) = \max\{(V_T/T)^{1/(\beta+2)}, 1/T^{1/(\beta+1)}\}$ when $\beta \geq 1$ and $\delta_V(\beta) = \max\{(V_T/T)^{1/3}, 1/\sqrt{T}\}$ when $0<\beta<1$.

Here we omit the details for the abrupt rotting (S_T).

Regret Analysis

We provide the theoretical results as follows.

Type	Regret upper bounds for $\beta \geq 1$	Regret upper bounds for $0<\beta<1$
V_T	$\widetilde{O}\left(\max\left\{V_T^{\frac{1}{\beta+2}}T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\right\}\right)$	$\widetilde{O}\left(\max\left\{V_T^{\frac{1}{3}}T^{\frac{2}{3}}, \sqrt{T}\right\}\right)$
S_T	$\widetilde{O}\left(\max\left\{S_T^{\frac{1}{\beta+1}}T^{\frac{\beta}{\beta+1}}, \bar{V}_T\right\}\right)$	$\widetilde{O}\left(\max\left\{\sqrt{S_T T}, \bar{V}_T\right\}\right)$

Type	Regret lower bounds for $\beta > 0$
V_T	$\Omega\left(\max\left\{V_T^{\frac{1}{\beta+2}}T^{\frac{\beta+1}{\beta+2}}, T^{\frac{\beta}{\beta+1}}\right\}\right)$
S_T	$\Omega\left(\max\left\{S_T^{\frac{1}{\beta+1}}T^{\frac{\beta}{\beta+1}}, \bar{V}_T\right\}\right)$

Experiments