A Change-Detection Based Framework for Piecewise-Stationary Multi-Armed Bandit Problem

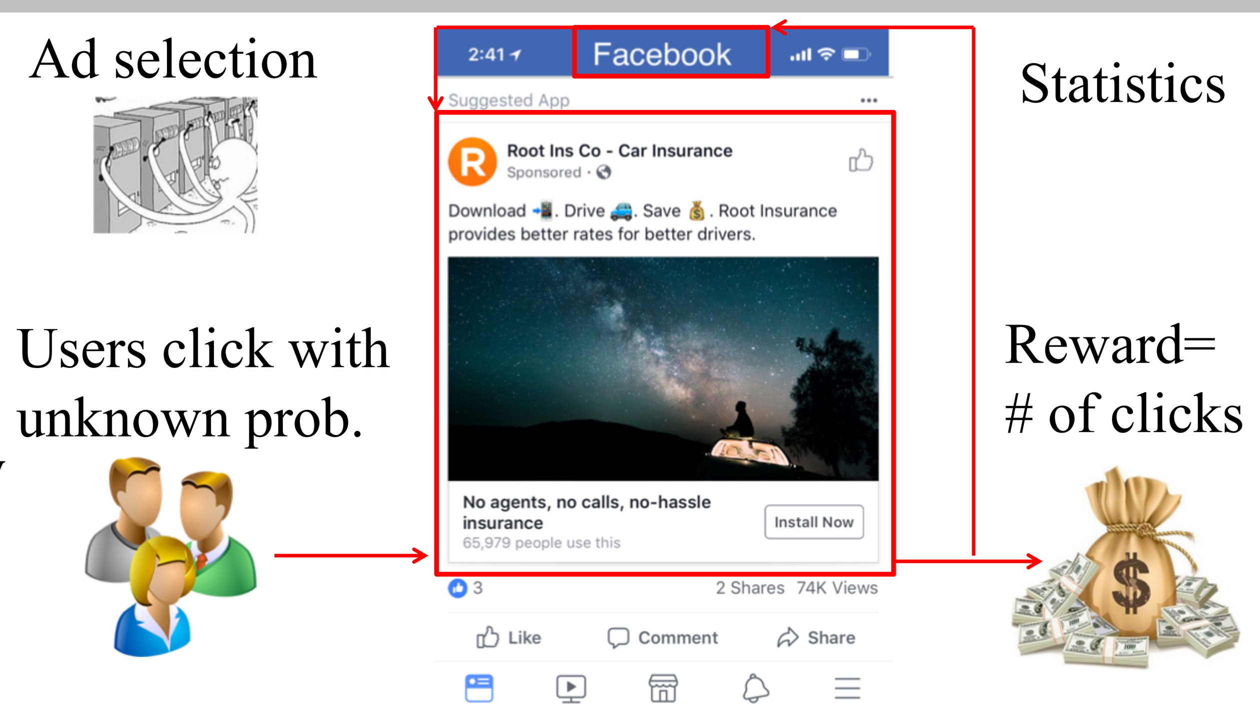
Fang Liu, Joohyun Lee and Ness Shroff

THE OHIO STATE UNIVERSITY

Introduction

Classical bandits:
- Slot machine;
- Unknown rewards;
- Stationary;
- O(logT) regret.

However, in real world problems, **non-stationary**
- User preference drift;
- Big event;
- Aging.



Existing methods:
- Passively adaptive policies - D-UCB, SW-UCB, Rexp3- with guarantee
- Actively adaptive policies - AdaptEvE, CTS - without guarantee

Model

Basic setting:
- Discrete time model of horizon T, there are K arms;
- At each time t, choosing an arm I_t returns a reward $X_t(I_t)$;
- The expectations, $\mu_t(i)$, may change over time;
- i^*_t:arm with the highest expected reward at time t.
- γ_T: number of change points up to time T.
- Regret: expected loss compared to the oracle that plays arm i^*_t each time.

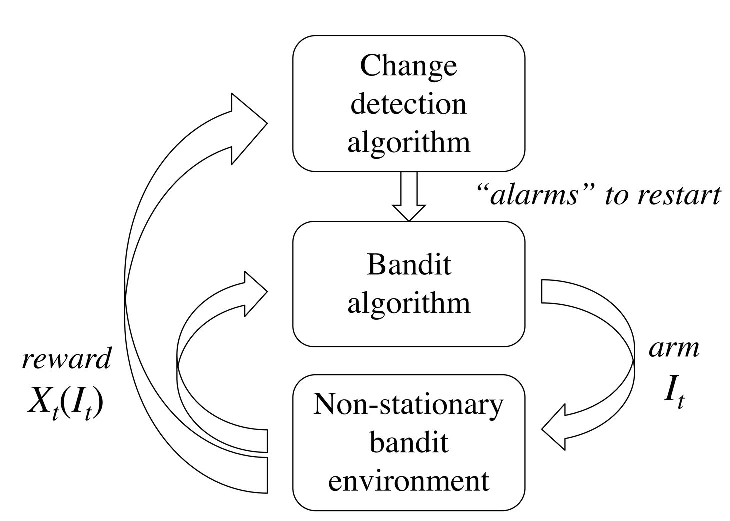
$$R_\pi(T) = \mathbb{E}\left[\sum_{t=1}^{T}(X_t(i_t^*) - X_t(I_t))\right]$$

Assumption 1: (piecewise stationarity) The shortest interval is larger than KM.
Assumption 2: (detectability) The expectation drift is no less than 3ϵ.
Assumption 3: Bernoulli rewards.

Algorithm

We propose change-detection based upper confidence bounds (CD-UCB).
- The change detection algorithm controls the restarting of UCB index;
- Mix the UCB decision with uniform sampling to feed CD algorithm.

We propose a tailored CUSUM algorithm for bandit problems.



Algorithm 1 CD-UCB

Require: T, α and an algorithm CD$(\cdot,\cdot)$
  Initialize $\tau_i = 1, \forall i$.
  for t **from** 1 **to** T **do**
    Update according to equations (3-5).
    Play arm I_t and observe $X_t(I_t)$.
    if CD$(I_t, X_t(I_t)) == 1$ **then**
      $\tau_{I_t} = t+1$; reset CD$(I_t,\cdot)$.
    end if
  end for

$$N_t(i) = \sum_{s=\tau_i}^{t} \mathbb{1}_{\{I_s=i\}}, \quad n_t = \sum_{i=1}^{K} N_t(i), \tag{3}$$

$$\bar{X}_t(i) = \sum_{s=\tau_i}^{t} \frac{X_s(i)}{N_t(i)} \mathbb{1}_{\{I_s=i\}}, \quad C_t(i) = \sqrt{\frac{\xi \log n_t}{N_t(i)}}, \tag{4}$$

Algorithm 2 Two-sided CUSUM

Require: parameters ϵ, M, h and $\{y_k\}_{k\geq 1}$
  Initialize $g_0^+ = 0$ and $g_0^- = 0$.
  for each k **do**
    Calculate s_k^- and s_k^+ according to (6).
    Update g_k^+ and g_k^- according to (7).
    if $g_k^+ \geq h$ or $g_k^- \geq h$ **then**
      Return 1
    end if
  end for

$$I_t = \begin{cases} \arg\max_{i\in\mathcal{K}} \left(\bar{X}_t(i) + C_t(i)\right), & \text{w.p. } 1-\alpha \\ i, & \forall i \in \mathcal{K}, \text{w.p. } \frac{\alpha}{K} \end{cases}. \tag{5}$$

$$(s_k^+, s_k^-) = (y_k - \hat{u}_0 - \epsilon, \hat{u}_0 - y_k - \epsilon)\mathbb{1}_{\{k>M\}}. \tag{6}$$

$$g_k^+ = \max(0, g_{k-1}^+ + s_k^+), \;\; g_k^- = \max(0, g_{k-1}^- + s_k^-). \tag{7}$$

Analysis

Theorem 1. (CD-UCB) *Let $\xi = 1$. Under Assumption 1, for any $\alpha \in [0,1)$ and any arm $i \in \{1,\dots,K\}$, the CD-UCB policy achieves,*

$$\mathbb{E}[\tilde{N}_T(i)] \leq (\gamma_T + \mathbb{E}[F]) \cdot \left(\frac{4\log T}{(\Delta_{\mu_T(i)})^2} + \frac{\pi^2}{3}\right) + \frac{\pi^2}{3} + \gamma_T \cdot \mathbb{E}[D] + \frac{\alpha T}{K}.$$

Corollary 1. (CD-UCB$|\alpha = 0$) *If $\alpha = 0$ and $\xi = 1$, then the regret of CD-UCB is*

$$R_{\pi^{\text{CD-UCB}}}(T) = O((\gamma_T + \mathbb{E}[F]) \cdot \log T + \gamma_T \cdot \mathbb{E}[D])).$$

Theorem 3. (CUSUM-UCB) *Let $\xi = 1$. Under Assumptions 1, 2 and 3, for any $\alpha \in (0,1)$ and any arm $i \in \{1,\dots,K\}$, the CUSUM-UCB policy achieves,*

$$\mathbb{E}[\tilde{N}_T(i)] \leq R_1 \cdot R_2 + \frac{\pi^2}{3} + \frac{\alpha T}{K},$$

for $R_1 = \gamma_T + \frac{2T}{(1-2\exp(-2\epsilon^2 M))\exp(C_1 h)}$, $R_2 = \frac{4\log T}{(\Delta_{\mu_T(i)})^2} + \frac{\pi^2}{3} + M + \frac{C_2(h+1)K}{\alpha}$.

Corollary 2. *Under the Assumptions 1, 2 and 3, if horizon T and the number of breakpoints γ_T are known in advance, then we can choose $h = \frac{1}{C_1}\log\frac{T}{\gamma_T}$ and $\alpha = K\sqrt{\frac{C_2\gamma_T}{C_1 T}\log\frac{T}{\gamma_T}}$ so that*

$$R_{\pi^{\text{CUSUM-UCB}}}(T) = O\left(\frac{\gamma_T \log T}{(\Delta_{\mu_T(i)})^2} + \sqrt{T\gamma_T \log\frac{T}{\gamma_T}}\right).$$

	Passively adaptive			*Actively adaptive*		
Policy	D-UCB (Kocsis and Szepesvári 2006)	SW-UCB (Garivier and Moulines 2008)	Rexp3 (Besbes, Gur, and Zeevi 2014)	Adapt-EvE (Hartland et al. 2007)	CUSUM-UCB	*lower bound* (Garivier and Moulines 2008)
Regret	$O(\sqrt{T\gamma_T}\log T)$	$O(\sqrt{T\gamma_T\log T})$	$O(V_T^{1/3}T^{2/3})$	Unknown	$O(\sqrt{T\gamma_T\log\frac{T}{\gamma_T}})$	$\Omega(\sqrt{T})$

Evaluation

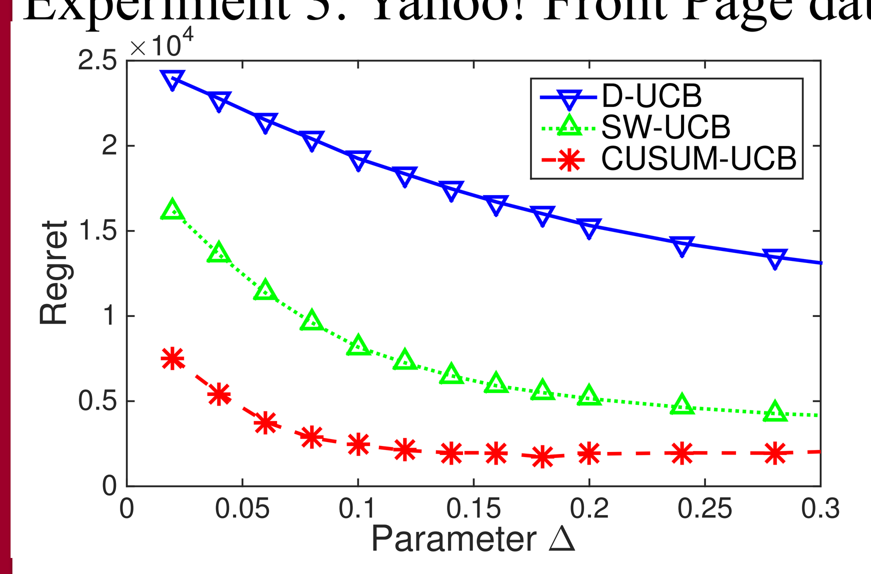
Experiment 1: Flipping environment. 2 Bernoulli arms with $\mu_t(1)$=0.5,

$$\mu_t(2) = \begin{cases} 0.5 - \Delta, & \frac{T}{3} \leq t \leq \frac{2T}{3} \\ 0.8, & \text{otherwise} \end{cases}.$$

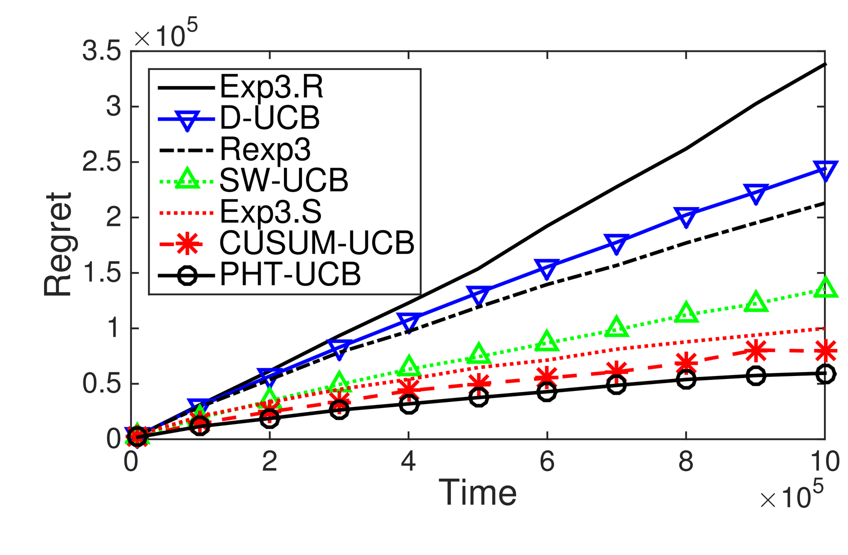
Experiment 2: Switching environment.

$$\mu_t(i) = \begin{cases} \mu_{t-1}(i), & \text{with probability } 1-\beta(t) \\ \mu \sim U[0,1], & \text{with probability } \beta(t) \end{cases}$$

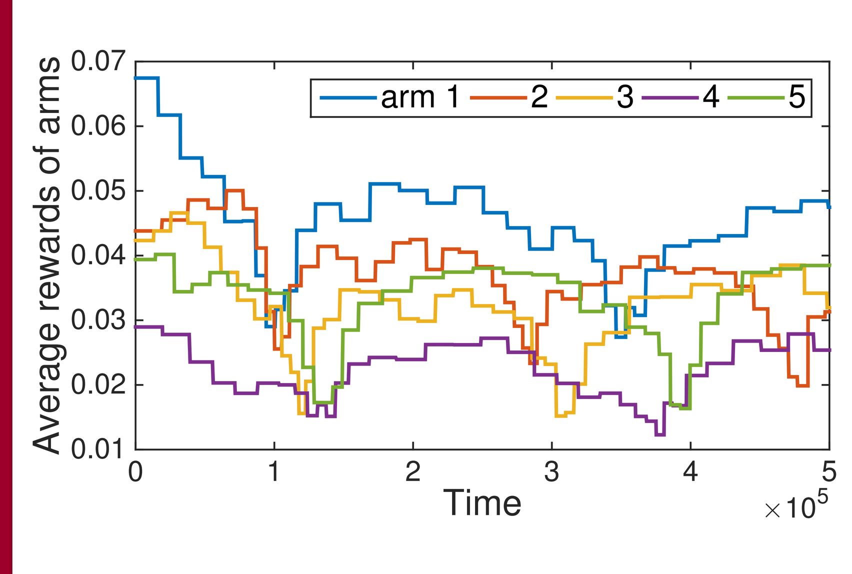
Experiment 3: Yahoo! Front Page dataset.



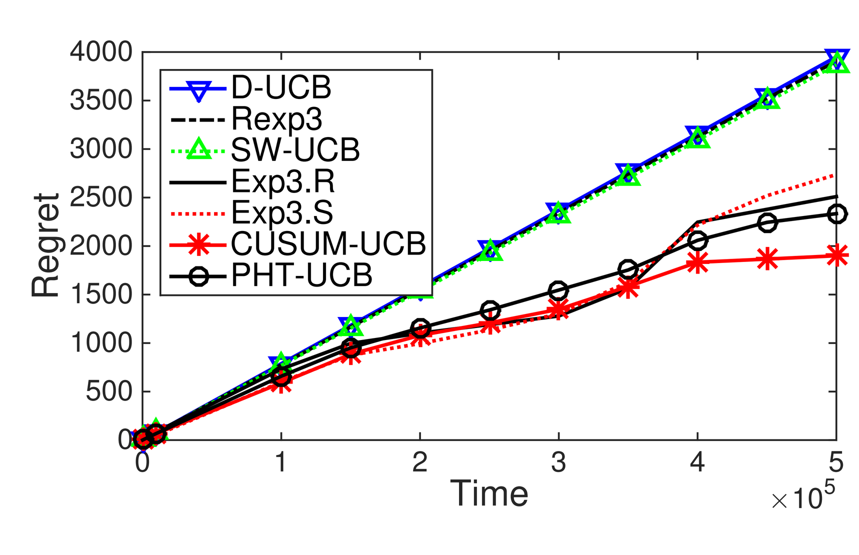
Flipping environment



Switching environment



Yahoo! ground truth



Yahoo! regret result