Piecewise-Stationary Combinatorial Bandits

Huozhi Zhou

University of Illinois at Urbana Champaign

hzhou35@illinois.edu

Sep. 14 2020

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Motivation

Multi-armed bandit (stochastic) can be defined by K different reward distributions $\{f_1, \ldots, f_k\}$ associated with different arms.

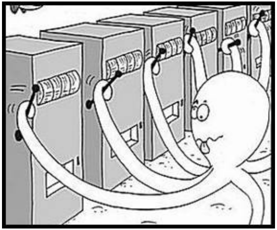
Goal: Identify the best arm.

Performance metric:

Regret: $\mathcal{R}(T) = T * \mu^* - \mathbb{E}[\sum_{i=1}^{T} \mu_t]$. (convergence rate)

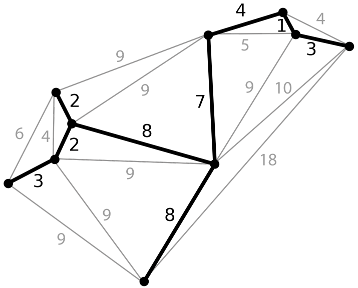
Sample Complexity: $n(\epsilon, \delta)$. (final performance)

Application: Advert placement; Resource allocation; Dynamic pricing.



Motivation

Many sequential decision making problems have combinatorial nature.
Network routing system optimization.



Goal: Minimize the expected total delay.

Motivation

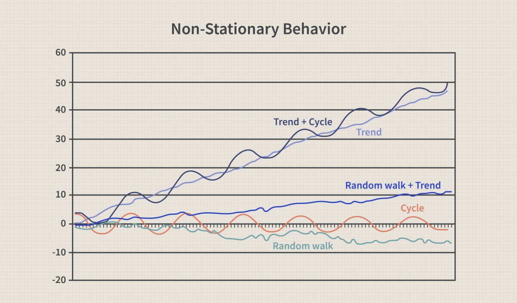
Top-m arm identification, consider advertise m products to the user.

Goal: Identify m most attractive products.

Motivation

What if the reward distributions are **non-stationary**?

- Network links might degrade over time.
- User's preference might change.



Motivation

- Many sequential decision making problems involve combinatorial action, and are in general non-stationary.
- Need a better model for quasi-stationary sequential decision making problems. Existing models are either too optimistic or too pessimistic.
- How good can we perform on this type of problem? Can we achieve optimal performance?

Piecewise-stationary combinatorial bandits

- We study two variants of piecewise-stationary combinatorial bandits, and develop efficient algorithms which achieve nearly order-optimal regret upper bound.
- Key idea of algorithm design is to balance uniform exploration and UCB-type exploration.
- By using randomized hard instance argument, we improve the minimax regret lower bound for piecewise-stationary bandits.

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Learning Protocol of CMAB

For $t = 1, \ldots, T$

- Based on the historical data, learner selects a superarm $S_t \in \mathcal{F}$.
- The reward of each base arm contained in the superarm is revealed to the learner $\{R_i(t) | i \in S_t\}$, as well as the reward of the superarm $R_{S_t}(t)$.

Piecewise-stationary CMAB: problem formulation

Piecewise-stationary CMAB $= (\mathcal{K}, \mathcal{F}, \mathcal{T}, \{f_{k,t}\}_{k \in \mathcal{K}, t \in \mathcal{T}}, r_{\boldsymbol{\mu}_t}(S_t))$.

- $\mathcal{K}$: set of base arms
- $\mathcal{F}$: set of super arms
- $\mathcal{T} = \{1, \dots, T\}$: time horizon
- $\{f_{k,t}\}_{k \in \mathcal{K}, t \in \mathcal{T}}$: collection of reward distributions of base arms throughout the time
- $r_{\boldsymbol{\mu}_t}(S_t)$: expected reward function

We assume reward distributions of base arms change in a piecewise manner, let $N = 1 + \sum_{t=1}^{T-1} \mathbb{I}\{\exists k \in \mathcal{K} \text{ s.t. } f_{k,t} \neq f_{k,t+1}\}$.
Goal: identify good super arm at each time step to achieve small regret

Piecewise-stationary CMAB: problem formulation

Assumption 1. (Monotoncity) Given two arbitrary mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, if $\mu_k \geq \mu'_k,\ \forall k \in \mathcal{K}$, then $r_{\boldsymbol{\mu}}(S) \geq r_{\boldsymbol{\mu}'}(S)$.
Assumption 2. [L-Lipschitz] Given two arbitrary mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, there exists an $L < \infty$ such that $\left| r_{\boldsymbol{\mu}}(S) - r_{\boldsymbol{\mu}'}(S) \right| \leq L \|\mathcal{P}_S(\boldsymbol{\mu} - \boldsymbol{\mu}')\|_2$, $\forall S \in \mathcal{F}$.
We assume access to an α-approximation oracle $\text{Oracle}_\alpha(\boldsymbol{\mu})$. Given a mean vector $\boldsymbol{\mu}$, $\text{Oracle}_\alpha(\boldsymbol{\mu})$ outputs an α-suboptimal super arm S such that $r_{\boldsymbol{\mu}}(S) \geq \alpha \max_{S \in \mathcal{F}} r_{\boldsymbol{\mu}}(S)$.
Performance metric: Expected α-approximation cumulative regret

$$\mathcal{R}(T) = \mathbb{E}\left[\alpha \sum_{t=1}^{T} \max_{S \in \mathcal{F}} r_{\boldsymbol{\mu}_t}(S) - \sum_{t=1}^{T} r_{\boldsymbol{\mu}_t}(S_t)\right],$$

One concrete example of the reward function

Consider top-m arm identification. In this case

$$r_{\mu_t}(S_t) = \sum_{i \in S_t} r_i(t)$$

which is the summation of rewards of all base arms contained in the super arm. We can verify that in this case r_{μ_t} is 1-Lipschitiz. The orcale can be realized by any sorting algorithm. In general, $r_{\mu_t}(\cdot)$ can be nonlinear with respect to the rewards of base arms.

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
    - Problem formulation
    - GLR change-point detector
    - `GLR-CUCB` algorithm
    - Regret upper bound
    - Minimax regret lower bound
    - Experimental results
- Piecewise-stationary Cascading Bandits
    - Problem formulation
    - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
    - Regret upper bound
    - Minimax regret lower bound
    - Experimental results

Piecewise-stationary CMAB: GLR change point detector

We use generalized likelihood (GLR) change point detector (Besson and Kaufmann, 2019) to monitor the change of base arms' reward distributions.

Algorithm 1: Sub-Bernoulli GLR Change-Point Detector: $\mathrm{GLR}(X_1, \cdots, X_n; \delta)$

Input: observations $X_1, \ldots, X_n$ and confidence level δ.
if $\sup_{s \in [1, n-1]} \quad [s \times kl(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}) + (n-s) \times kl(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n})] \geq \beta(n, \delta)$ **then**
  Return **True**
end
else
  Return **False**
end

Advantage: Almost parameter-free.

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Piecewise-stationary CMAB: GLR-CUCB algorithm

GLR-CUCB, runs in three phases:

1. For p fraction of the time, the algorithm play a superarm by uniform exploration. For the rest of the time, we play the superarm according to the α-approximation oracle (use UCB indices as input).

2. Once a superarm is played, the algorithm update the UCB indices of all based arm contained in the superarm:

$\mathrm{UCB}(k) \leftarrow \frac{1}{n_\ell} \sum_{n=1}^{n_k} Z_{k,n} + \sqrt{\frac{3\log(t-\tau)}{2n_k}}$.

3. At the end of each round, run GLR change-point detector on all base arms contained in the played superarm. If at least one base arm has changed, restart the UCB indices of all base arms.

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Piecewise-stationary CMAB: suboptimal gaps

The set of bad super arms with respect to the ith piecewise-stationary segment is defined as:

$$\mathcal{S}_B^i = \{S | r_{\boldsymbol{\mu}^i}(S) \leq \alpha \max_{\tilde{S} \in \mathcal{F}} r_{\boldsymbol{\mu}^i}(\tilde{S})\}$$

The *suboptimality gaps* in the ith stationary segment are:

$$\begin{aligned}
\Delta_{\text{opt}}^{\min,i} &= \alpha \max_{\tilde{S} \in \mathcal{F}} r_{\boldsymbol{\mu}^i}(\tilde{S}) - \max\{r_{\boldsymbol{\mu}^i}(S) | S \in \mathcal{S}_B^i\}, \\
\Delta_{\text{opt}}^{\max,i} &= \alpha \max_{\tilde{S} \in \mathcal{F}} r_{\boldsymbol{\mu}^i}(\tilde{S}) - \min\{r_{\boldsymbol{\mu}^i}(S) | S \in \mathcal{S}_B^i\}.
\end{aligned}$$

Denote the largest gap at change-point ν_i as

$$\Delta_{\text{change}}^i = \max_{k \in \mathcal{K}} \left| \mu_k^{i+1} - \mu_k^i \right|, \forall 1 \leq i \leq N-1.$$

Piecewise-stationary CMAB: regret upper bound

Assumption 3. Define $d_i = d_i(p,\delta) = \left\lceil \{4K/p\left(\Delta^i_{\text{change}}\right)^2\}\beta(T,\delta) + \frac{K}{p} \right\rceil$ and assume $\nu_i - \nu_{i-1} \geq 2\max\{d_i, d_{i-1}\}$, $\forall i = 1, \ldots, N-1$, where $\nu_N - \nu_{N-1} \geq 2d_{N-1}$.

Remark: The length of each piecewise-stationary segment is $\Omega(\sqrt{T \log T})$.

Theorem (Zhou et al. 2020)

The expected α-approximation cumulative regret of `GLR-CUCB` *with exploration probability p and confidence level δ satisfies*

$$\mathcal{R}(T) \leq \sum_{i=1}^{N} \widetilde{C}_i + \Delta_{\text{opt}}^{max} Tp + \sum_{i=1}^{N-1} \Delta_{\text{opt}}^{max,i+1} d_i + 3NT\Delta_{\text{opt}}^{max} K\delta,$$

where $\widetilde{C}_i = \left(6L^2K^2 \log T / \left(\Delta_{\text{opt}}^{min,i}\right)^2 + \pi^2/6 + K\right) \Delta_{\text{opt}}^{max,i}$.

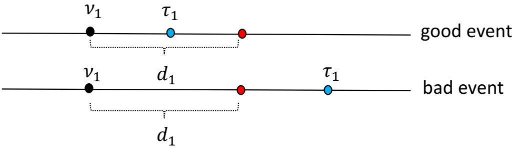
Piecewise-stationary CMAB: proof sketch

High-level idea: Recursive regret decomposition.

Step 1: For stationary case, we have

$$\mathcal{R}(T) \leq \underbrace{\Delta_{\text{opt}}^{\max,1} T \mathbb{P}(\tau_1 \leq T)}_{\text{false alarm}} + \underbrace{\Delta_{\text{opt}}^{\max,1} Tp}_{\text{uniform exploration}} + \underbrace{\widetilde{C}_1}_{\text{UCB exploration}}.$$

Step 2: Event decomposition.



Define good events $\{C^{(i)}\}_{i=1}^{N-1}$

$$\mathcal{C}^{(i)} = \{\forall j \leq i, \tau_j \in \{\nu_j + 1, \cdots, \nu_j + d_j\}\}, i \in [N-1].$$

Piecewise-stationary CMAB: proof sketch

Define $F_i = \{\tau_i > \nu_i\}$ and $D_i = \{\tau_i \le \nu_i + d_i\}$, $\forall 1 \le i \le N-1$. By definition, $\mathcal{C}^{(i)} = F_1 \cap D_1 \cap \cdots \cap F_i \cap D_i$.

Step 3: Regret decomposition.

First decompose with respect to F_1.

$$\begin{aligned}
\mathcal{R}(T) = \mathbb{E}\left[R(T)\right] &= \mathbb{E}\left[R(T)\mathbb{I}\{F_1\}\right] + \mathbb{E}\left[R(T)\mathbb{I}\{\overline{F}_1\}\right] \\
&\le \mathbb{E}\left[R(T)\mathbb{I}\{F_1\}\right] + T\Delta_{\text{opt}}^{\max}\mathbb{P}(\overline{F}_1) \\
&\le \mathbb{E}\left[R(\nu_1)\mathbb{I}\{F_1\}\right] + \mathbb{E}\left[R(T-\nu_1)\right] + T\Delta_{\text{opt}}^{\max}K\delta \\
&\le \widetilde{C}_1 + \Delta_{\text{opt}}^{\max,1}\nu_1 p + \mathbb{E}\left[R(T-\nu_1)\right] + T\Delta_{\text{opt}}^{\max}K\delta,
\end{aligned}$$

Then decompose $\mathbb{E}[R(T-\nu_1)]$ with respect to $\mathcal{C}^{(1)}$. Repeat the above procedure for $\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(N-1)}$, we can obtain the desired bound.

Piecewise-stationary CMAB: regret upper bound

Corollary (Zhou et al. 2020)

Let $\Delta_{\text{change}}^{min} = \min_{i\in[N-1]} \Delta_{\text{change}}^{i}$*, we have*

1. *(N is known) Choosing* $\delta = \frac{1}{T}$, $p = \sqrt{\frac{NK \log T}{T}}$*, gives*
$$\mathcal{R}(T) = \mathcal{O}\left(\frac{NK^2 \log T \Delta_{\text{opt}}^{max}}{\left(\Delta_{\text{opt}}^{min}\right)^2} + \frac{\sqrt{NKT \log T}\Delta_{\text{opt}}^{max}}{\left(\Delta_{\text{change}}^{min}\right)^2}\right);$$
2. *(N is unknown) Choosing* $\delta = \frac{1}{T}$, $p = \sqrt{\frac{K \log T}{T}}$*, gives*
$$\mathcal{R}(T) = \mathcal{O}\left(\frac{NK^2 \log T \Delta_{\text{opt}}^{max}}{\left(\Delta_{\text{opt}}^{min}\right)^2} + \frac{N\sqrt{KT \log T}\Delta_{\text{opt}}^{max}}{\left(\Delta_{\text{change}}^{min}\right)^2}\right).$$

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Piecewise-stationary CMAB: minimax regret lower bound

Theorem (Zhou et al. 2020)

If $K \geq 3$ and $T \geq M_1 N \frac{(K-1)^2}{K}$, then the worst-case regret for any policy is lower bounded by

$$\mathcal{R}(T) \geq M_2 \sqrt{NKT},$$

where $M_1 = 1/\log\frac{4}{3}$, $M_2 = 1/24\sqrt{\log\frac{4}{3}}$.

Proof sketch:

1. Construct randomized hard instance. $\mu_i^{i^*} \sim \mathrm{Bern}(\frac{1}{2} + \epsilon)$, $\forall k \in \mathcal{K} \setminus i^*, \mu_i^k \sim \mathrm{Bern}(\frac{1}{2})$. $(i+1)^* | i^* \sim \mathrm{uniform}(\mathcal{K} \setminus i^*)$
2. By change of measure technique, one can show that this ensemble of hard instances incurs regret at least $\Omega(\sqrt{NKT})$.
3. There exists at least one instance incurs regret on the order of $\Omega(\sqrt{NTK})$.

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
    - Problem formulation
    - GLR change-point detector
    - `GLR-CUCB` algorithm
    - Regret upper bound
    - Minimax regret lower bound
    - Experimental results
- Piecewise-stationary Cascading Bandits
    - Problem formulation
    - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
    - Regret upper bound
    - Minimax regret lower bound
    - Experimental results

Piecewise-statinoary CMAB: experimental results

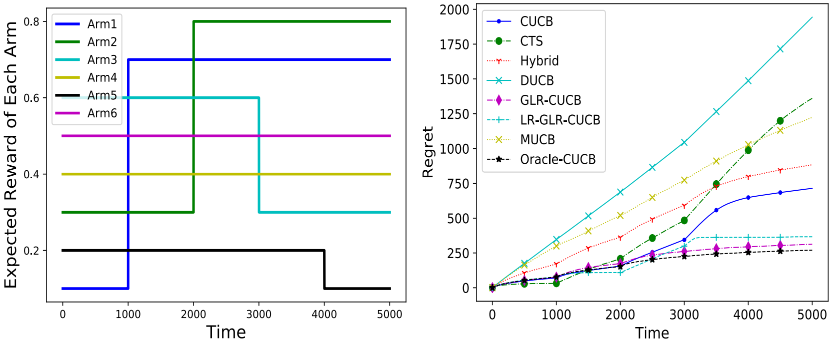


Figure: Experiment on synthetic dataset. Left: reward distribution of base arms. Right: expected accumulative regret.

Piecewise-statinoary CMAB: experimental results

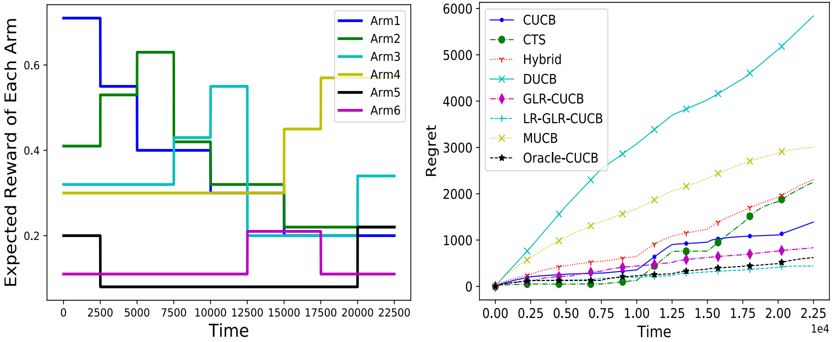
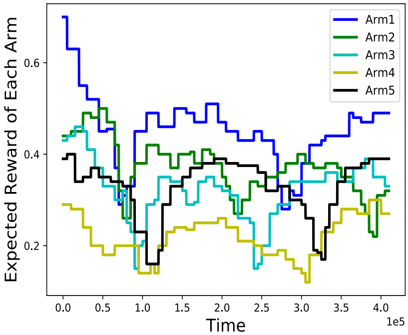


Figure: Yahoo! 1 experiment. Left: reward distribution of base arms. Right: expected accumulative regret.



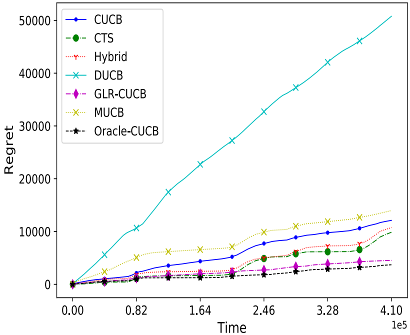


Figure: Yahoo! 2 experiment. Left: reward distribution of base arms. Right: expected accumulative regret.

Piecewise-stationary CMAB: experimental results

	CUCB	CTS	Hybrid	DUCB	GLR-CUCB
Synthetic Dataset	241.08	351.12	278.82	14.08	37.96
Yahoo! Experiment 1	510.20	513.41	826.01	25.44	62.76
Yahoo! Experiment 2	563.54	562.24	1189.17	158.27	517.13

Table: Standard deviations of all algorithms for experiments on synthetic and Yahoo! datasets

	LR-GLR-CUCB	MUCB	Oracle-CUCB
Synthetic Dataset	73.30	171.45	25.54
Yahoo! Experiment 1	63.37	202.44	35.08
Yahoo! Experiment 2	496.73	1427.28	160.76

Table: Standard deviations of all algorithms for experiments on synthetic and Yahoo! datasets

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Learning protocal of cascading bandit

For $t = 1, \dots, T$

- Given historical data, the learner selects K out of L items to recommend to the user.
- The learner observes a partial feedback of his decision, $\arg\min_k 1 \le k \le K : Z_{a_{k,t}, t} = 1$, the first item clicked by the user/no click.

Piecewise-stationary cascading bandit: problem formulation

Piecewise-stationary cascading bandit (CB)=$(\mathcal{L}, \mathcal{T}, \{f_{\ell,t}\}_{\ell\in\mathcal{L},t\in\mathcal{T}}, K)$.

- $\mathcal{L}$: Ground set containing L items (e.g., web pages or advertisements).
- $\mathcal{T} = \{1, \ldots, T\}$: Set of time steps.
- $\{f_{\ell,t}\}_{\ell\in\mathcal{L},t\in\mathcal{T}}$: Pmfs of items in $\mathcal{L}$ at all time steps.
- K: Number of items recommended by the learner to the user.

Learner receives partial feedback at time t, given by

$$F_t = \begin{cases} \emptyset, & \text{if no click,} \\ \arg\min_k \{1 \le k \le K : Z_{a_{k,t},t} = 1\}, & \text{otherwise.} \end{cases}$$

Goal of learner: Identify top-K items with highest clicked probabilities.

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} R\left(\mathcal{A}_t, \mathbf{w}_t, \mathbf{Z}_t\right)\right],$$

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

`GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithm run in three phases:

1. For p fraction of the time, the algorithm select K items by uniform sampling. For the rest of the time, the algorithm select K items with highest UCB/KL-UCB indices.

2. Update the statistics of K selected items.

$$\mathrm{UCB}(\ell) = \hat{\mathbf{w}}(\ell) + \sqrt{\frac{3\log(t-\tau)}{2n_\ell}},$$

$$\mathrm{UCB}_{\mathrm{KL}}(\ell) = \max\{q \in [\hat{\mathbf{w}}(\ell), 1] : n_\ell \times \mathrm{KL}(\hat{\mathbf{w}}(\ell), q) \leq g(t-\tau)\}.$$

3. At the end of each round, run GLR change-point detector on selected items at this round. If at least one item's click probability has changed, restart the UCB indices/KL-UCB indices of all items.

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Piecewise-stationary cascading bandit: regret upper bound

Theorem (Wang et al. 2019)

`GLRT-CascadeUCB` *guarantees*

$$\mathcal{R}(T) \leq \underbrace{\sum_{i=1}^{N} \widetilde{C}_i}_{(a)} + \underbrace{Tp}_{(b)} + \underbrace{\sum_{i=1}^{N-1} d_i}_{(c)} + \underbrace{3NTL\delta}_{(d)},$$

where $\widetilde{C}_i = \sum_{\ell=K+1}^{L} \frac{12}{\Delta^{i}_{s_i(\ell), s_i(K)}} \log T + \frac{\pi^2}{3} L.$

Piecewise-stationary cascading bandit: regret upper bound

Theorem (Wang et al. 2019)

`GLRT-CascadeKL-UCB` *guarantees*

$$\mathcal{R}(T) \leq \underbrace{T(N-1)(L+1)\delta}_{(a)} + \underbrace{Tp}_{(b)} + \underbrace{\sum_{i=1}^{N-1} d_i}_{(c)} + \underbrace{NK \log\log T + \sum_{i=0}^{N-1} \widetilde{D}_i}_{(d)},$$

where $\widetilde{D}_i$ *is a term depending on* $\log T$ *and the suboptimal gaps.*

Piecewise-stationary cascading bandit: regret upper bound

Corollary (Wang et al. 2019)

The regret of `GLRT-CascadeUCB` *is established by choosing* $\delta = \frac{1}{T}$ *and* $p = \sqrt{\frac{NL \log T}{T}}$*:*

$$\mathcal{R}(T) = \mathcal{O}\left(\frac{N(L-K)\log T}{\Delta_{\text{opt}}^{\text{min}}} + \frac{\sqrt{NLT \log T}}{\left(\Delta_{\text{change}}^{\text{min}}\right)^2}\right). \tag{1}$$

Choosing the same δ *and* p *,* `GLRT-CascadeKL-UCB` *has same order of regret upper bound as* (1)*.*

remark: the order of the regret upper bound is the same as `GLR-CUCB`, which implies that the dominant factor is the change in distribution.

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Piecewise-stationary cascading bandit: minimax regret lower bound

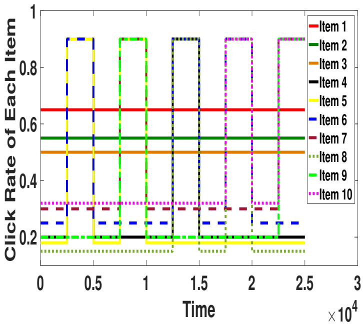
Theorem (Wang et al. 2019)

If $L \geq 3$ *and* $T \geq MN\frac{(L-1)^2}{L}$*, then for any policy, the worst-case regret is at least* $\Omega(\sqrt{NLT})$*, where* $M = 1/\log\frac{4}{3}$*, and* $\Omega(\cdot)$ *notation hides a constant factor that is independent of* N*,* L*, and* T*.*

Outline

- Motivation
- Piecewise-stationary Combinatorial Semi-Bandits
  - Problem formulation
  - GLR change-point detector
  - `GLR-CUCB` algorithm
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results
- Piecewise-stationary Cascading Bandits
  - Problem formulation
  - `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms
  - Regret upper bound
  - Minimax regret lower bound
  - Experimental results

Piecewise-stationary cascading bandit: experimental results



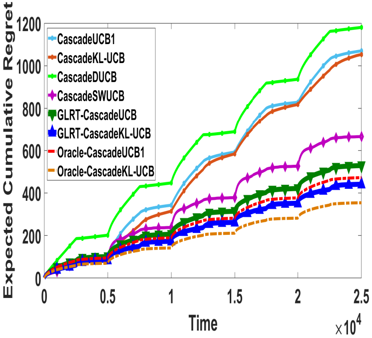


Figure: Synthetic experiment. Left: reward distributions. Right: cumulative regret.

Piecewise-stationary cascading bandit: experimental results

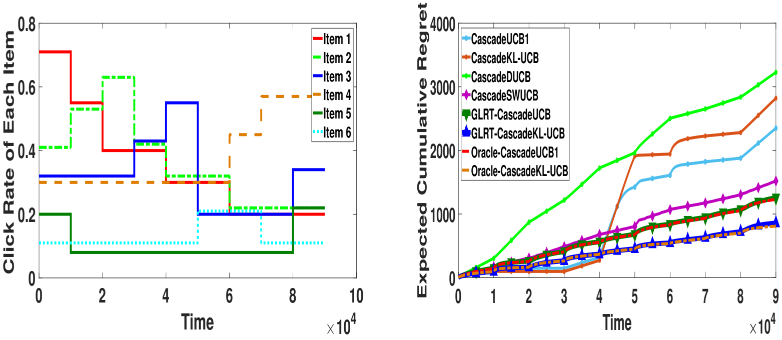


Figure: Experiment on Yahoo! dataset. Left: reward distributions. Right: cumulative regret.

Piecewise-stationary cascading bandit: experimental results

Table: Means and standard deviations of the T-step regrets.

	CascadeUCB1	CascadeKL-UCB	CascadeDUCB
Synthetic Dataset	1069.77 ± 87.09	1053.25 ± 111.67	1180.30 ± 20.22
Yahoo! Experiment	2349.29 ± 312.71	2820.16 ± 256.74	3226.97 ± 39.37
	CascadeSWUCB	GLRT-CascadeUCB	GLRT-CascadeKL-UCB
Synthetic Dataset	664.84 ± 29.81	$\mathbf{527.93 \pm 25.20}$	$\mathbf{440.93 \pm 45.54}$
Yahoo! Experiment	1519.56 ± 52.23	$\mathbf{1235.21 \pm 54.59}$	$\mathbf{856.77 \pm 67.16}$
	Oracle-CascadeUCB1	Oracle-CascadeKL-UC	
Synthetic Dataset	472.25 ± 17.65	353.86 ± 19.59	
Yahoo! Experiment	1230.17 ± 45.24	808.84 ± 47.97	

Conclusion

- We develop the first efficient algorithm for piecewise-stationary combinatorial semi-bandits, `GLR−CUCB`, which achieves $\mathcal{O}(\sqrt{NKT \log T})$ regret.
- We improve minimax regret lower bound ($\Omega(\sqrt{NKT})$) for piecewise-stationary combinatorial semi-bandits, which indicates `GLR−CUCB` is nearly order-optimal within poly-logarithm factors.
- We develop better algorithms for piecewise-stationary cascading bandits and tighten the minimax regret lower bound.
- Future work includes design time-unaware algorithms for piecewise-stationary bandits, incorporate contextual information, etc.

References

Besson, Lilian and Emilie Kaufmann. "The Generalized Likelihood Ratio Test meets klUCB: an Improved Algorithm for Piece-Wise Non-Stationary Bandits". In: *arXiv preprint arXiv:1902.01575* (2019).

Wang, Lingda, Huozhi Zhou, Bingcong Li, Lav R. Varshney, and Zhizhen Zhao. "Be Aware of Non-Stationarity: Nearly Optimal Algorithms for Piecewise-Stationary Cascading Bandits". In: *arXiv preprint arxiv:1909.05886* (2019).

Zhou, Huozhi, Lingda Wang, Lav R Varshney, and Ee-Peng Lim. "A Near-Optimal Change-Detection Based Algorithm for Piecewise-Stationary Combinatorial Semi-Bandits". In: *Proc. 34st AAAI Conf. Artif. Intell (AAAI'20)*. 2020.

Q & A