ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414506

NEAR-OPTIMAL ALGORITHMS FOR PIECEWISE-STATIONARY CASCADING BANDITS

Lingda Wang[*,1], *Huozhi Zhou*[*,1], *Bingcong Li*[†], *Lav R. Varshney*[*], *Zhizhen Zhao*[*]

[*]University of Illinois at Urbana-Champaign [†]University of Minnesota - Twin Cities

ABSTRACT

Cascading bandit (CB) is a popular model for web search and online advertising. However, the stationary CB model may be too simple to cope with real-world problems, where user preferences may change over time. Considering piecewise-stationary environments, two efficient algorithms, `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB`, are developed. Comparing with existing works, the proposed algorithms: i) are free of change-point-dependent information for choosing parameters; ii) have fewer tuning parameters; iii) improve regret upper bounds. We also show that the proposed algorithms are optimal up to logarithm terms by deriving a minimax lower bound $\Omega(\sqrt{NLT})$ for piecewise-stationary CB. The efficiency of the proposed algorithms is validated through numerical tests on a real-world benchmark dataset.

***Index Terms*—** Online Learning, Cascading Bandits, Non-stationary Bandits, Change-point Detection.

1. INTRODUCTION

Online recommendation and web search are of significant importance in the modern economy. Based on a user's browsing history, these systems strive to maximize satisfaction and minimize regret by presenting the user with a list of items (e.g., web pages and advertisements) that meet her/his preference. Such a scenario can be modeled via cascading bandits (CB) [1], where an agent aims to identify the K most attractive items out of total L items contained in the ground set.

CB can be viewed as multi-armed bandits (MAB) tailored for cascade model (CM) [2] that depicts a user's online behavior. Existing works on CB [1] and MAB [3–6] can be categorized according to whether stationary or non-stationary environment is studied. In stationary environments, the attraction (or reward) distributions do not evolve over time. On the other hand, non-stationary environments are prevalent in real-world applications, as user's preference is time-varying [7, 8]. The most common non-stationary environments include adversarial [9], piecewise-stationary [10], and slow-varying [11].

In this paper, we focus on the piecewise-stationary environment, where the user's preference remains stationary over a few time slots, named *piecewise-stationary segments*, but can shift abruptly at some unknown slots, called *change-points*. To address the piecewise-stationary MAB, two types of approaches have been proposed in the literature: *passively adaptive approaches* [10–12] and *actively adaptive approaches* [13–16]. Passively adaptive approaches ignore when a change-point occurs. For active adaptive approaches, a change-point detection algorithm such as CUSUM [14, 17] and Page Hinkley Test (PHT) [14, 18] is included. Within the area of piecewise-stationary CB (PS-CB), only passively adaptive approaches have been studied [19]. In this context, we introduce the generalized likelihood ratio test (GLRT) [15, 20] for actively adaptive CB algorithms. In particular, we develop two GLRT based algorithms `GLRT-Cascade-UCB` and `GLRT-CascadeKL-UCB` to enhance both theoretical and practical effectiveness for PS-CB. The merits of this paper are summarized as follows

1. **Practically oriented.** The proposed GLRT based algorithms are more practical than previous works [13, 14], since: i) no change-point-dependent parameter is required by GLRT; ii) fewer tuning parameters are required.
2. **Tighter regret bounds.** Both algorithms are shown to have regret bounds $\mathcal{O}(\sqrt{NLT\log T})$, where L is the number of items and T is the number of time slots. Our regret bound tightens those of [19] by a factor of $\sqrt{L}$ and $\sqrt{L\log T}$, respectively.
3. **Lower-bound matching.** We establish that the minimax regret lower bound for PS-CB is $\Omega(\sqrt{NLT})$. Such a lower bound: i) implies the proposed algorithms are optimal up to a logarithm factor; ii) is the first to characterize dependence on N, L, and T for PS-CB.
4. **Numerically attractive.** Numerical experiments on a real-world benchmark dataset reveal the merits of proposed algorithms over state-of-the-art approaches.

2. PROBLEM FORMULATION

2.1. Cascade Model and Cascading Bandits

CB [1], as a learning variant of CM, depicts the interaction between the agent and the user on T time slots. CM [2] explains the user's behavior in a specific time slot t.

In CM, the user is presented with a K-item ranked list $\mathcal{A}_t := (a_{1,t}, \ldots, a_{K,t}) \in \Pi_K(\mathcal{L})$ from $\mathcal{L}$ at time slot t, where $\mathcal{L} := \{1, 2, \ldots, L\}$ is a ground set containing L items (e.g., web pages or advertisements), and $\Pi_K(\mathcal{L})$ is the set of all K-permutations of $\mathcal{L}$. CM can be parameterized by the attraction probability vector $\mathbf{w}_t = [\mathbf{w}_t(1), \ldots, \mathbf{w}_t(L)]^\top \in [0,1]^L$.

[1]indicates equal contributions. This work has been supported by the IBM-Illinois Center for Cognitive Computing Systems Research, and the Alfred P. Sloan Foundation.

The user browses the list $\mathcal{A}_t$ from the first item a_1 in order, and each item a_k attracts the user to click it with probability $\mathbf{w}_t(a_k)$. The user will stop the process after clicking the first attractive item. In particular, when an item $a_{k,t}$ is clicked, it means that i) items from $a_{1,t}$ to $a_{k-1,t}$ are not attractive to the user; and ii) items $a_{k+1,t}$ to $a_{K,t}$ are not browsed so whether they are attractive to the user is unknown.

Building upon CM, a CB problem can be described by a tuple $(\mathcal{L}, \mathcal{T}, \{f_{\ell,t}\}_{\ell\in\mathcal{L},t\in\mathcal{T}}, K)$, where $\mathcal{T} := \{1, 2, \ldots, T\}$ collects all T time slots. Whether the user is attracted by item ℓ at time slot t is denoted by a Bernoulli random variable $Z_{\ell,t}$, whose pmf is $f_{\ell,t}$. As convention, $Z_{\ell,t} = 1$ indicates item ℓ is attractive to the user. We also denote $\mathbf{Z}_t := \{Z_{\ell,t}\}_{\ell\in\mathcal{L}}$ as all the attraction variables. Clearly, the $\{f_{\ell,t}\}_{\ell\in\mathcal{L},t\in\mathcal{T}}$ are parameterized by the attraction probability vectors $\{\mathbf{w}_t\}_{t\in\mathcal{T}}$, which are unknown to the agent. Since CB is for stationary environments, $\mathbf{w}_t$ is time-invariant, and can be simplified as $\mathbf{w}$. CB poses a mild assumption on $\{f_{\ell,t}\}_{\ell\in\mathcal{L},t\in\mathcal{T}}$.

Assumption 1. *The attraction distributions $\{f_{\ell,t}\}_{\ell\in\mathcal{L},t\in\mathcal{T}}$ are independent both across items and time slots.*

Per slot t, the agent recommends a list of K items $\mathcal{A}_t$ to the user based on the feedback up to time slot $t-1$. The feedback at time slot t refers to the index of the clicked item, given by

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

After the user browses $\mathcal{A}_t$ follows the protocol described by CM, the agent observes the feedback F_t. Along with F_t is a zero-one reward indicating whether there is a click

$$r\left(\mathcal{A}_t, \mathbf{Z}_t\right) = 1 - \prod_{k=1}^{K}\left(1 - Z_{a_{k,t},t}\right), \tag{1}$$

where $r\left(\mathcal{A}_t, \mathbf{Z}_t\right) = 0$ if $F_t = \emptyset$. Then, this process proceeds to time slot $t+1$. The goal of the agent is to maximize the expected cumulative reward over the whole time horizon $\mathcal{T}$. Noticing that $Z_{\ell,t}$s are independent, the expected reward at time slot t can be computed as $\mathbb{E}\left[r\left(\mathcal{A}_t, \mathbf{Z}_t\right)\right] = r\left(\mathcal{A}_t, \mathbf{w}\right)$. The optimal list $\mathcal{A}^*$ remains the same for all time slots, which is the list containing the K most attractive items.

2.2. Piecewise-Stationary Cascading Bandits

The stationarity assumption on CB limits its applicability for real-world applications, as users tend to change their preferences as time goes on [7]. This fact leads to PS-CB. Consider a PS-CB problem with N segments, where the attraction probabilities of items remain identical per segment. Mathematically, N can be written as

$$N = 1 + \sum_{t=1}^{T-1} \mathbb{I}\{\exists \ell \in \mathcal{L} \text{ s.t. } f_{\ell,t} \neq f_{\ell,t+1}\}, \tag{2}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, and a change-point is the time slot t that satisfies $\exists \ell \in \mathcal{L}$ s.t. $f_{\ell,t} \neq f_{\ell,t+1}$. These change-points are denoted by $\nu_1, \ldots, \nu_{N-1}$ in a chronological manner. Specifically, $\nu_0 = 0$ and $\nu_N = T$ are introduced for consistency. For the ith piecewise-stationary segment $t \in [\nu_{i-1}+1, \nu_i]$, f_ℓ^i and $\mathbf{w}^i(\ell)$ denote the attraction distribution and the expected attraction of item ℓ, respectively, which are again unknown to the agent. Attraction probability vector $\mathbf{w}^i = [\mathbf{w}^i(1), \ldots, \mathbf{w}^i(L)]^\top$ collects $\mathbf{w}^i(\ell)$s.

In a PS-CB problem, the agent interacts with its users in the same manner of CB. The agent's policy is evaluated by its expected cumulative reward, or its expected cumulative regret:

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} R\left(\mathcal{A}_t, \mathbf{w}_t, \mathbf{Z}_t\right)\right], \tag{3}$$

where the expectation $\mathbb{E}[\cdot]$ is taken with respect to a sequence of $\mathbf{Z}_t$ and the corresponding $\mathcal{A}_t$. Here, $R(\mathcal{A}_t, \mathbf{w}_t, \mathbf{Z}_t) = r(\mathcal{A}_t^*, \mathbf{w}_t) - r(\mathcal{A}_t, \mathbf{Z}_t)$ is the regret at time slot t with

$$\mathcal{A}_t^* = \arg\max_{\mathcal{A}_t \in \Pi_K(\mathcal{L})} r\left(\mathcal{A}_t, \mathbf{w}_t\right)$$

being the optimal list that maximizes the expected reward.

3. ALGORITHMS

3.1. Generalized Likelihood Ratio Test

Our adaptive change-point detection relies on the GLRT summarized in Algorithm 1. Compared with existing change-point detection methods with provable guarantees [13, 14], advantages of GLRT are twofold: i) *Fewer tuning parameters*. GLRT only requires one parameter, while CUSUM [14] and CMSW [13] have three and two parameters, respectively. ii) *Less prior knowledge needed*. GLRT does not require the information on the smallest magnitude among the change-points, which is essential for CUSUM.

Algorithm 1 GLRT Change-Point Detector.

Require: observations $X_1, \ldots, X_n$ and confidence level δ
1: Compute the GLR statistic $\mathrm{GLR}(n)$ according to (4) and the threshold $\beta(n,\delta)$ according to (4)
2: Return True **if** $\mathrm{GLR}(n) \ge \beta(n,\delta)$ **else** False.

Next, GLRT is formally introduced. Suppose we have a sequence of Bernoulli random variables $\{X_t\}_{t=1}^n$ and aim to determine if a change-point exists. GLRT adopts GLR statistic as the judgement, which is

$$\mathrm{GLR}(n) = \sup_{s\in[1,n-1]} [s \times \mathrm{KL}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{KL}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right)].$$

Here, $\hat{\mu}_{s:s'}$ is the empirical mean of observations from X_s to $X_{s'}$, and $\mathrm{KL}(x, y)$ is the Kullback–Leibler (KL) divergence. By comparing $\mathrm{GLR}(n)$ in (4) with the threshold $\beta(t,\delta)$, one can decide whether a change-point appears, where

$$\beta(t,\delta) = 2\mathcal{G}\left(\log(3t\sqrt{t}/\delta)/2\right) + 6\log(1+\log t), \tag{4}$$

and $\mathcal{G}(\cdot)$ has the same definition as that in (13) of [21]. The choice of δ influences the sensitivity of the GLRT. For example, a larger δ makes the GLRT response faster to a change-point, but increases the probability of false alarm.

3.2. The GLRT Based CB Algorithms

The proposed algorithms, `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB`, are presented in Algorithm 2. On a high level, three phases comprise the proposed algorithms. *Phase 1*: The forced uniform exploration to ensure that sufficient

samples are gathered for all items to perform the GLRT detection. *Phase 2*: The UCB-based exploration (UCB or KL-UCB) to learn the optimal list on each piecewise-stationary segment. *Phase 3*: The GLRT change-point detection to monitor if global restart should be triggered.

By recommending the list $\mathcal{A}_t$ and observing the user's feedback F_t (line 9), we update the statistics (line 11) and perform the GLRT detection (line 12). If a change-point is detected, we set $n_\ell = 0$ for all $\ell \in \mathcal{L}$, and $\tau = t$ (line 13). Finally, the UCB indices of each item are computed as follows (line 18),

$$\text{UCB}(\ell) = \hat{\mathbf{w}}(\ell) + \sqrt{3\log(t-\tau)/(2n_\ell)}, \tag{5}$$

$$\text{UCB}_{\text{KL}}(\ell) = \max\{q \in [\hat{\mathbf{w}}(\ell), 1] : n_\ell \times \text{KL}(\hat{\mathbf{w}}(\ell), q) \le g(t-\tau)\}, \tag{6}$$

where $g(t) = \log t + 3\log\log t$.

Algorithm 2 The Proposed Algorithms

Require: The time horizon $\mathcal{T}$, the ground set $\mathcal{L}$, K, exploration probability $p > 0$, and confidence level $\delta > 0$

1: **Initialization:** $\tau \leftarrow 0$ and $n_\ell \leftarrow 0, \forall \ell \in \mathcal{L}$
2: **for all** $t = 1, 2, \ldots, T$ **do**
3: $a \leftarrow (t-\tau) \mod \lfloor \frac{L}{p} \rfloor$
4: **if** $a \le L$ **then**
5: Choose $\mathcal{A}_t$ such that $a_{1,t} \leftarrow a$ and $a_{2,t}, \ldots, a_{K,t}$ are chosen uniformly at random
6: **else**
7: Compute $\mathcal{A}_t = \arg\max_{\mathcal{A} \in \Pi_K(\mathcal{L})} r(\mathcal{A}, \text{UCB or UCB}_{\text{KL}})$
8: **end if**
9: Recommend the list $\mathcal{A}_t$ to user, and observe feedback F_t
10: **for all** $k = 1, \ldots, F_t$ **do**
11: $\ell \leftarrow a_{k,t}$, $n_\ell \leftarrow n_\ell + 1$, $X_{\ell,n_\ell} \leftarrow \mathbb{I}\{F_t = k\}$ and $\hat{\mathbf{w}}(\ell) = \frac{1}{n_\ell}\sum_{n=1}^{n_\ell} X_{\ell,n}$
12: **if** GLRT$(X_{\ell,1}, \ldots, X_{\ell,n_\ell}; \delta)$ = True **then**
13: $n_\ell \leftarrow 0$, $\forall \ell \in \mathcal{L}$, and $\tau \leftarrow t$
14: **end if**
15: **end for**
16: **for** $\ell = 1, \cdots, L$ **do**
17: **if** $n_\ell \neq 0$ **then**
18: Compute UCB(ℓ) according to (5) for `GLRT-CascadeUCB` or UCB$_{\text{KL}}(\ell)$ according to (6) for `GLRT-CascadeKL-UCB`
19: **end if**
20: **end for**
21: **end for**

Besides the time horizon $\mathcal{T}$, the ground set $\mathcal{L}$, the number of items in list K, the proposed algorithms only require two parameters p and δ as inputs. The probability p is used to control the portion of uniform exploration in Phase 1. The confidence level δ is the parameter required by GLRT, and the choice of δ will be discussed in Section 4. In Algorithm 2, we denote the last detection time as τ. From slot τ to current slot, let n_l denote the number of observations for ℓth item, and $\hat{\mathbf{w}}(\ell)$ its corresponding sample mean. The algorithm determines whether to perform a uniform or UCB-based exploration depending on line 4, which ensures the fraction of time slots performing the uniform exploration is about p. If the uniform exploration is triggered, the first item in the recommended list $\mathcal{A}_t$ will be item a in Line 3, and the remaining items in the list are chosen at random (line 5), which ensures item a will be observed by the user. If UCB-based exploration is adopted at time slot t, the algorithm chooses K items (line 7) with K largest UCB indices,

4. THEORETICAL RESULTS

Without loss of generality, for the ith piecewise-stationary segment, the ground set $\mathcal{L}$ is first sorted in decreasing order according to attraction probabilities, that is $\mathbf{w}^i(s_i(1)) \ge \cdots \ge \mathbf{w}^i(s_i(L)), \forall s_i(\ell) \in \mathcal{L}$. The optimal list at ith segment is thus all the permutations of the set $\mathcal{A}_i^* = \{s_i(1), \ldots, s_i(K)\}$. The item ℓ^* is optimal if $\ell^* \in \mathcal{A}_i^*$, otherwise an item ℓ is called suboptimal. The gap between the attraction probabilities of ℓ and ℓ^* at ith segment is defined as: $\Delta^i_{\ell,\ell^*} = \mathbf{w}^i(\ell^*) - \mathbf{w}^i(\ell)$. Similarly, the largest amplitude change among items at change-point ν_i is defined as $\Delta^i_{\text{change}} = \max_{\ell\in\mathcal{L}} \left|\mathbf{w}^{i+1}(\ell) - \mathbf{w}^i(\ell)\right|$, with $\Delta^0_{\text{change}} = \max_{\ell\in\mathcal{L}} \left|\mathbf{w}^1(\ell)\right|$. We have the following assumption for the theoretical analysis.

Assumption 2. *Define* $d_i = \lceil \frac{4L\beta(T,\delta)}{p(\Delta^i_{\text{change}})^2} + \frac{L}{p} \rceil$ *and assume* $\nu_i - \nu_{i-1} \ge 2\max\{d_i, d_{i-1}\}$, $\forall i = 1, \ldots, N-1$.

Note that Assumption 2 is standard in a piecewise-stationary environment, and similar assumptions can be found in other actively adaptive approaches [13–15] as well. Assumption 2 guarantees that with high probability all the change-points are detected within the interval $[\nu_i + 1, \nu_i + d_i]$, which is equivalent to saying all change-points are detected correctly and quickly.

4.1. Regret Upper Bound for `GLRT-CascadeUCB`

The regret of `GLRT-CascadeUCB` is as follows.

Theorem 1. *Suppose that Assumptions 1 and 2 are satisfied,* `GLRT-CascadeUCB` *guarantees*

$$\mathcal{R}(T) \le \sum_{i=1}^{N} \widetilde{C}_i + Tp + \sum_{i=1}^{N-1} d_i + 3NTL\delta,$$

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

Proof. See supplemental material online at [22]. □

Corollary 1 follows directly from Theorem 1.

Corollary 1. *Let* $\Delta^{\min}_{\text{change}} = \min_{i\le N-1} \Delta^i_{\text{change}}$ *denote the smallest magnitude of any change-point on any item, and* $\Delta^{\min}_{\text{opt}} = \min_{i\le N} \Delta^i_{s_i(K+1),s_i(K)}$ *be the smallest magnitude of a suboptimal gap on any one of the stationary segments. The regret of* `GLRT-CascadeUCB` *is established by choosing* $\delta = 1/T$ *and* $p = \sqrt{NL\log T/T}$*:*

$$\mathcal{R}(T) = \mathcal{O}\left(\frac{N(L-K)\log T}{\Delta^{\min}_{\text{opt}}} + \frac{\sqrt{NLT\log T}}{\left(\Delta^{\min}_{\text{change}}\right)^2}\right). \tag{7}$$

Proof. See supplemental material online at [22]. □

As T becomes larger, the regret is dominated by the cost of the change-point detection component, implying the regret is $\mathcal{O}(\sqrt{NLT\log T}/(\Delta^{\min}_{\text{change}})^2)$.

4.2. Regret Upper Bound for GLRT-CascadeKL-UCB

The regret of GLRT-CascadeKL-UCB is as follows.

Theorem 2. *Suppose that Assumptions 1 and 2 are satisfied,* GLRT-CascadeKL-UCB *guarantees*

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

where $\widetilde{D}_i$ is a term depending on $\log T$ and the suboptimal gaps. Detailed expression can be found in (10) in the supplemental material online at [22].

Proof. See supplemental material online at [22]. □

Corollary 2. *Choosing the same δ and p as in Corollary 1,* GLRT-CascadeKL-UCB *has the same regret as (7).*

4.3. Minimax Regret Lower Bound

In this subsection, we derive a minimax regret lower bound for PS-CB, which is tighter than $\Omega(\sqrt{T})$ proved in [19].

Theorem 3. *If $L \geq 3$ and $T \geq N(L-1)^2/(L\log(4/3))$, then for any policy, the worst-case regret is at least $\Omega(\sqrt{NLT})$.*

Proof. See supplemental material online at [22]. □

This lower bound is the first characterization involving N, L, and T. And it indicates our proposed algorithms are nearly order-optimal within a logarithm factor $\sqrt{\log T}$.

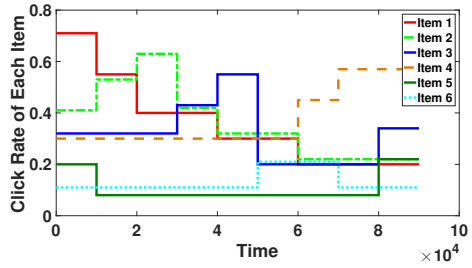
4.4. Discussion

The regrets of GLRT-CascadeUCB and GLRT-CascadeKL-UCB can be upper bounded by $\mathcal{R}(T) = \mathcal{O}\left(\sqrt{NLT\log T}\right)$. Note that compared to CUSUM in [14] and CMSW in [13], the tuning parameters are fewer and does not require the smallest magnitude among the change-points $\Delta^{\min}_{\text{change}}$ as shown in Corollary 1. Moreover, parameter δ and p follow simple rules as shown in Corollary 1, while complicated parameter tuning steps are required in CUSUM and CMSW. The regrets of proposed algorithms are improved over state-of-the-art algorithms in [19] either in the dependence on L or both L and T, as their upper bounds are $\mathcal{O}(L\sqrt{NT}\log T)$ and $\mathcal{O}(L\sqrt{NT\log T})$, respectively.

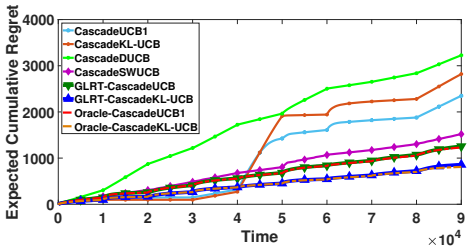
5. NUMERICAL TESTS

In this section, we carry out numerical tests on the Yahoo! benchmark dataset[1] designed for the evaluation of bandit algorithms to validate the effectiveness of proposed algorithms. Four baseline algorithms are chosen for comparison, where CascadeUCB1 [1] and CascadeKL-UCB [1] are near-optimal algorithms to handle stationary CB; while CascadeDUCB [19] and Cascade-SWUCB [19] cope with piecewise-stationary CB through a passively adaptive manner. In addition, two oracle algorithms, Oracle-CascadeUCB1 and Oracle-Cascade-KL-UCB, that have access to change-point times are also selected for comparison. In particular, the oracle algorithms restart when a change-point occur. Based on the theoretical analysis by [19], we choose $\xi = 0.5$, $\gamma = 1 - 0.25/\sqrt{T}$ for CascadeDUCB and choose $\tau = 2\sqrt{T\log T}$ for CascadeSWUCB. For GLRT-CascadeUCB and GLRT-Ca-scadeKL-UCB, we set $\delta = 1/T$ and $p = 0.1\sqrt{N\log T/T}$.

We pre-process the dataset by adopting the same method as [13], where $L = 6$, $K = 2$ and $N = 9$. To make the experiment nontrivial, several modifications are applied to the dataset: i) the click rate of each item is enlarged by 10 times; ii) the time horizon is reduced to $T = 90000$, which is shown in Figure 1a. Figure 1b presents the cumulative regrets of all algorithms by averaging 100 trials, which shows the regrets of our proposed algorithms are just slightly above the oracle algorithms and significantly outperform other algorithms.



(a) Click rate



(b) Expected cumulative reret

Fig. 1: (a) Click rate of each item of Yahoo! dataset with $T = 90000$, $L = 6$ and $N = 9$, and (b) Expected cumulative regrets of different algorithms on Yahoo! dataset.

6. CONCLUSION

Two new actively adaptive algorithms for piecewise-stationary cascading bandit are developed in this work, which achieve the same near-optimal regret upper bound on the order of $\mathcal{O}\left(\sqrt{NLT\log T}\right)$. This matches our minimax regret lower bound up to a $\sqrt{\log T}$ factor. Compared with state-of-the-art algorithms that adopt passively adaptive approach such as CascadeSWUCB and CascadeDUCB, our new regret upper bounds are reduced by $\mathcal{O}(\sqrt{L})$ and $\mathcal{O}(\sqrt{L\log T})$ respectively. Numerical tests on Yahoo! dataset show the improved efficiency of the proposed algorithms. Several interesting questions are still left open for future work. One challenging problem lies in whether the $\sqrt{\log T}$ gap in time steps T between regret upper bound and lower bound can be closed. In addition, we are also interested in extending the single click models to multiple clicks models in future work.

[1] https://webscope.sandbox.yahoo.com

7. REFERENCES

[1] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan, "Cascading bandits: Learning to rank in the cascade model," in *Proc. 32th Int. Conf. Mach. Learn. (ICML 2015)*, 2015, pp. 767–776.

[2] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey, "An experimental comparison of click position-bias models," in *Proc. 1st ACM Int. Conf. Web Search Data Min. (WSDM'08)*. ACM, 2008, pp. 87–94.

[3] Tze Leung Lai and Herbert Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, 2002.

[5] Bingcong Li, Tianyi Chen, and Georgios B Giannakis, "Bandit online learning with unknown delays," in *Proc. 22th Int. Conf. Artif. Intell. Stat. (AISTATS 2019)*, 2019, pp. 993–1002.

[6] Lingda Wang, Bingcong Li, Huozhi Zhou, Georgios B Giannakis, Lav R Varshney, and Zhizhen Zhao, "Adversarial linear contextual bandits with graph-structured side observations," *arXiv preprint arXiv:2012.05756*, 2020.

[7] Rolf Jagerman, Ilya Markov, and Maarten de Rijke, "When people change their mind: Off-policy evaluation in non-stationary recommendation environments," in *Proc. 12th ACM Int. Conf. Web Search Data Min. (WSDM'19)*. ACM, 2019, pp. 447–455.

[8] Jia Yuan Yu and Shie Mannor, "Piecewise-stationary bandit problems with side observations," in *Proc. 26th Int. Conf. Mach. Learn. (ICML 2009)*. ACM, 2009, pp. 1177–1184.

[9] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[10] Aurélien Garivier and Eric Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. 22th Int. Conf. Algorithmic Learning Theory (ALT'11)*. 2011, pp. 174–188, Springer.

[11] Omar Besbes, Yonatan Gur, and Assaf Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Proc. 24th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS'14)*, 2014, pp. 199–207.

[12] Lai Wei and Vaibhav Srivatsva, "On abruptly-changing and slowly-varying multiarmed bandit problems," in *Proc. Am. Contr. Conf. (ACC 2018)*. IEEE, 2018, pp. 6291–6296.

[13] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie, "Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit," in *Proc. 22nd Int. Conf. Artif. Intell. Stat. (AISTATS 2019)*, 2019, pp. 418–427.

[14] Fang Liu, Joohyun Lee, and Ness Shroff, "A change-detection based framework for piecewise-stationary multi-armed bandit problem," in *Proc. 32nd AAAI Conf. Artif. Intell (AAAI'18).*, 2018.

[15] Lilian Besson and Emilie Kaufmann, "The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits," *arXiv preprint arXiv:1902.01575*, 2019.

[16] Huozhi Zhou, Lingda Wang, Lav R Varshney, and Ee-Peng Lim, "A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits," in *Proc. 34th AAAI Conf. Artif. Intell*, 2020, vol. 34, pp. 6933–6940.

[17] Ewan S Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[18] David V Hinkley, "Inference about the change-point from cumulative sum tests," *Biometrika*, vol. 58, no. 3, pp. 509–523, 1971.

[19] Chang Li and Maarten de Rijke, "Cascading non-stationary bandits: Online learning to rank in the non-stationary cascade model," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI 2019)*, 2019, pp. 2859–2865.

[20] Alan Willsky and H Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Trans. Autom. Control*, vol. 21, no. 1, pp. 108–112, 1976.

[21] Emilie Kaufmann and Wouter Koolen, "Mixture martingales revisited with applications to sequential tests and confidence intervals," *arXiv preprint arXiv:1811.11419*, 2018.

[22] Lingda Wang, Huozhi Zhou, Bingcong Li, Lav R Varshney, and Zhizhen Zhao, "Nearly optimal algorithms for piecewise-stationary cascading bandits," *arXiv preprint arXiv:1909.05886*, 2019.