Cascading Non-stationary Bandits: Online Learning to Rank in the Non-stationary Cascade Model

Li, C.; de Rijke, M.

Publication date
2019

Document Version
Author accepted manuscript

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Cascading Non-Stationary Bandits: Online Learning to Rank in the Non-Stationary Cascade Model

Chang Li, Maarten de Rijke
University of Amsterdam
{c.li, derijke}@uva.nl

Abstract
Non-stationarity appears in many online applications such as web search and advertising. In this paper, we study the online learning to rank problem in a non-stationary environment where user preferences change abruptly at an unknown moment in time. We consider the problem of identifying the $K$ most attractive items and propose cascading non-stationary bandits, an online learning variant of the cascading model, where a user browses a ranked list from top to bottom and clicks on the first attractive item. We propose two algorithms for solving this non-stationary problem: CascadeDUCB and CascadeSWUCB. We analyze their performance and derive gap-dependent upper bounds on the $n$-step regret of these algorithms. We also establish a lower bound on the regret for cascading non-stationary bandits and show that both algorithms match the lower bound up to a logarithmic factor. Finally, we evaluate their performance on a real-world web search click dataset.

1 Introduction
Learning to rank LTR [Liu, 2009] is a combination of machine learning and information retrieval. It is a core problem in many applications, such as web search and recommendation [Liu, 2009; Zoghi et al., 2017]. The goal of LTR is to rank items, e.g., documents, and show the top $K$ items to a user. Traditional LTR algorithms are supervised, offline algorithms; they learn rankers from human annotated data [Qin et al., 2010] and/or users’ historical interactions [Joachims, 2002]. Every day billions of users interact with modern search engines and leave a trail of interactions. It is feasible and important to design online algorithms that directly learn from such user clicks to help improve users’ online experience. Indeed, recent studies show that even well-trained production rankers can be optimized by using users’ online interactions, such as clicks [Zoghi et al., 2016].

Generally, interaction data is noisy [Joachims, 2002], which gives rise to the well-known exploration vs. exploitation dilemma. Multi-armed bandit (MAB) [Auer et al., 2002a] algorithms have been designed to balance exploration and exploitation. Based on MABs, many online LTR algorithms have been published [Radlinski et al., 2008; Kveton et al., 2015; Katariya et al., 2016; Lagrèe et al., 2016; Zoghi et al., 2017; Li et al., 2019]. These algorithms address the exploration vs. exploitation dilemma in an elegant way and aim to maximize user satisfaction in a stationary environment where users do not change their preferences over time. Moreover, they often come with regret bounds.

Despite the success of the algorithms mentioned above in the stationary case, they may have linear regret in a non-stationary environment where users may change their preferences abruptly at any unknown moment in time. Non-stationarity widely exists in real-world application domains, such as search engines and recommender systems [Yu and Mannor, 2009; Pereira et al., 2018; Wu et al., 2018; Jagerman et al., 2019]. Particularly, we consider abruptly changing environments where user preferences remain constant in certain time periods, named epochs, but change occurs abruptly at unknown moments called breakpoints. The abrupt changes in user preferences give rise to a new challenge of balancing “remembering” and “forgetting” [Besbes et al., 2014]: the more past observations an algorithm retains the higher the risk of making a biased estimator, while the fewer observations retained the higher stochastic error it has on the estimates of the user preferences.

In this paper, we propose cascading non-stationary bandits, an online variant of the cascade model (CM) [Craswell et al., 2008] with the goal of identifying the $K$ most attractive items in a non-stationary environment. CM is a widely-used model of user click behavior [Chuklin et al., 2015; Zoghi et al., 2017]. In CM, a user browses the ranked list from top to bottom and clicks the first attractive item. The items ranked above the first clicked item are browsed but not attractive since they are not clicked. The items ranked below the first clicked item are not browsed since the user stops browsing the ranked list after a click. Although CM is a simple model, it effectively explains user behavior [Kveton et al., 2015].

Our key technical contributions in this paper are: (1) We formalize a non-stationary online learning to rank (OLTR) problem as cascading non-stationary bandits. (2) We propose two algorithms, CascadeDUCB and CascadeSWUCB, for solving it. They are motivated by discounted UCB (DUCB) and sliding window UCB (SWUCB), respectively [Garivier and Moulines, 2011]. CascadeDUCB balances “remembering” and “forgetting” by using a discounting factor of past observations, and
CascadeSWUCB balances the two by using statistics inside a fixed-size sliding window. (3) We derive gap-dependent upper bounds on the regret of the proposed algorithms. (4) We derive a lower bound on the regret of cascading non-stationary bandits. We show that the upper bounds match this lower bound up to a logarithmic factor. (5) We evaluate the performance of CascadeSWUCB and CascadeDUCB empirically on a real-world web search click dataset.

2 Background

We define the learning problem at the core of this paper in Assumption 1.

We refer readers to [Chuklin et al., 2015] for an introduction to click models. Briefly, a click model models a user’s interaction behavior with the search system. The user is presented with a $K$-item ranked list $R$. Then the user browses the list $R$ and clicks items that potentially attract him or her. Many click models have been proposed and each models a certain aspect of interaction behavior. We can parameterize a click model by attraction probabilities $\alpha \in [0,1]^K$ and a click model assumes:

**Assumption 1.** The attraction probability $\alpha(a)$ only depends on item $a$ and is independent of other items.

CM is a widely-used click model [Craswell et al., 2008; Zoghi et al., 2017]. In the CM, a user browses the ranked list $R$ from the first item $R(1)$ to the last item $R(K)$, which is called the cascading assumption. After the user browses an item $R(i)$, he or she clicks on $R(i)$ with attraction probability $\alpha(R(i))$, and then stops browsing the remaining items. Thus, the examination probability of item $R(j)$ equals the probability of clicking any item in the list: $1 - \prod_{i=1}^{K}(1 - \alpha(R(i)))$. The expected number of clicks equals the probability of clicking any item in the list: $1 - \prod_{i=1}^{K}(1 - \alpha(R(i)))$. Note that the reward does not depend on the order in $R$, and thus, in the CM, the goal of ranking is to find the $K$ most attractive items.

The CM accepts at most one click in each search session. It cannot explain scenarios where a user may click multiple items. The CM has been extended in different ways to capture multi-click cases [Chapelle and Zhang, 2009; Guo et al., 2009]. Nevertheless, CM is still the fundamental click model and its historical click data reasonably well. Thus, in this paper, we focus on the CM and in the next section we introduce an online variant of CM, called cascading bandits.

2.2 Cascading bandits

Cascading bandits (CB) is defined by a tuple $B = (D, P, K)$, where $D = [L]$ is the set of candidate items, $K \leq L$ is the number of positions, $P \in \{0,1\}^K$ is a distribution over binary attractions.

In CB, at time $t$, a learning agent builds a ranked list $R_t \in \Pi_K(D)$ that depends on the historical observations up to $t$ and shows it to the user. $A_t \in \{0,1\}^K$ is defined as the attraction indicator, which is drawn from $P$ and $A_t(R_t(i))$ is the attraction indicator of item $R_t(i)$. The user examines $R_t$ from $R_t(1)$ to $R_t(K)$ and clicks the first attractive item. Since a CM allows at most one click each time, a random variable $c_t$ is used to indicate the position of the clicked item, i.e., $c_t = \arg\min_{i \in [K]} I\{A_t(R_t(i))\}$. If there is no attractive item, the user will not click, and we set $c_t = K+1$ to indicate this case. Specifically, if $c_t \leq K$, the user clicks an item, otherwise, the user does not click anything. After the click or browsing the last item in $R_t$, the user leaves the search session. The click feedback $c_t$ is then observed by the learning agent. Because of the cascading assumption, the agent knows that items ranked above position $c_t$ are observed. The reward at time $t$ is defined by the number of clicks:

$$ r(R_t, A_t) = 1 - \prod_{i=1}^{K}(1 - A_t(R_t(i))) .$$  

(1)

Under Assumption 1, the attraction indicators of each item in $D$ are independently distributed. Moreover, cascading bandits make another assumption.

**Assumption 2.** The attraction indicators are distributed as:

$$ P(A) = \prod_{a \in D} P_a(A(a)) ,$$  

(2)

where $P_a$ is a Bernoulli distribution with a mean of $\alpha(a)$. Under Assumption 1 and 2, the attraction indicator of item $a$ at time $t$ $A_t(a)$ is drawn independently from other items. Thus, the expectation of reward of the ranked list at time $t$ can be computed as $\mathbb{E}[r(R_t, A_t)] = r(R_t, \alpha)$. And the goal of the agent is to maximize the expected number of clicks in $n$ steps.

Cascading bandits are designed for a stationary environment, where the attraction probability $P$ remains constant. However, in real-world applications, users change their preferences constantly [Jagerman et al., 2019], which is called a non-stationary environment, and learning algorithms proposed for cascading bandits, e.g., CascadeKL-UCB and CascadeUB1 [Kveton et al., 2015], may have linear regret in this setting. In the next section, we propose cascading non-stationary bandits, the first non-stationary variant of cascading bandits, and then propose two algorithms for solving this problem.

3 Cascading Non-Stationary Bandits

We first define our non-stationary online learning setup, and then we propose two algorithms learning in this setup.

3.1 Problem setup

The learning problem we study is called cascading non-stationary bandits, a variant of CB. We define it by a tuple
Algorithm 1: UCB-type algorithm for Cascading non-stationary bandits.

1: Input: discounted factor $\gamma$ or sliding window size $\tau$
2: // Initialization
3: $\forall a \in D: N_{i}(a) = 0$
4: $\forall a \in D: X_{i}(a) = 0$
5: for $t = 1, 2, \ldots, n$ do
6: for $a \in D$ do
7: // Compute UCBs
8: $U_{t}(a) \leftarrow \text{Eq. 5} \quad \text{(CascadeDUCB)}$
9: $\gamma \leftarrow \text{Eq. 7} \quad \text{(CascadeSWUCB)}$
10: // Recommend top $K$ items and receive clicks
11: $R_{t} \leftarrow \arg \max_{R \in \Pi_{K}} r_{t}(R, U_{t})$
12: Show $R_{t}$ and receive clicks $c_{t} \in \{1, \ldots, K + 1\}$
13: // Update statistics
14: if CascadeDUCB then
15: $\forall a \in D: N_{i}(a) = \gamma N_{i-1}(a)$
16: $\forall a \in D: X_{i}(a) = \gamma X_{i-1}(a)$
17: else
18: // for CascadeSWUCB
19: $\forall a \in D: N_{i}(a) = \sum_{s=t}^{t+1} 1\{a \in R_{s}\}$
20: $\forall a \in D: X_{i}(a) = \sum_{s=t}^{t+1} 1\{R_{s}^{-1}(a) = c_{s}\}$
21: for $i = 1, \ldots, \min\{c_{t}, K\}$ do
22: $a \leftarrow R_{t}(i)$
23: $N_{i}(a) = N_{i}(a) + 1$
24: $X_{i}(a) = X_{i}(a) + 1\{i = c_{t}\}$

CascadeDUCB is inspired by DUCB and CascadeSWUCB is inspired by SWUCB [Garivier and Moulines, 2011]. We summarize the pseudocode of both algorithms in Algorithm 1.

CascadeDUCB and CascadeSWUCB learn in a similar pattern. They differ in the way they estimate the Upper Confidence Bound (UCB) $U_{t}(R_{t}(i))$ of the attraction probability of item $R_{t}(i)$ at time $t$, as discussed later in this section. After estimating the UCBs (line 8), both algorithms construct $R_{t}$ by including the top $K$ most relevant items by UCB. Since the order of top $K$ items only affects the observation but does not affect the payoff of $R_{t}$, we construct $R_{t}$ as follows:

$$R_{t} = \arg \max_{R \in \Pi_{K}} r_{t}(R, U_{t}).$$  

After receiving the user’s click feedback $c_{t}$, both algorithms update their statistics (line 13–24). We use $N_{t}(i)$ and $X_{t}(i)$ to indicate the number of items $i$ that have been observed and clicked up to $t$ step, respectively.

To tackle the challenge of non-stationarity, CascadeDUCB penalizes old observations with a discount factor $\gamma \in (0, 1)$. Specifically, each of the previous statistics is discounted by $\gamma$ (line 15–16). The UCB of item $a$ is estimated as:

$$U_{t}(a) = \bar{\alpha}_{t}(\gamma, a) + c_{t}(\gamma, a),$$

where $\bar{\alpha}_{t}(\gamma, a) = \frac{X_{t}(a)}{N_{t}(a)}$ is the average of discounted attraction indicators of item $i$ and

$$c_{t}(\gamma, a) = 2 \sqrt{\frac{\ln N_{t}(\gamma)}{N_{t}(a)}}$$

is the confidence interval around $\bar{\alpha}_{t}(i)$ at time $t$. Here, we compute $\bar{\alpha}_{t}(\gamma) = \frac{1 - e^{\gamma N_{t}(\gamma)}}{1 - e^{\gamma N_{t}(\gamma)}}$ as the discounted time horizon. As shown in [Garivier and Moulines, 2011], $\alpha_{t}(a) \in [\bar{\alpha}_{t}(\gamma, a) - c_{t}(\gamma, a), \bar{\alpha}_{t}(\gamma, a) + c_{t}(\gamma, a)]$ holds with high probability.

As to CascadeSWUCB, it estimates UCBs by observations inside a sliding window with size $\tau$. Specifically, it only considers the observations in the previous $\tau$ steps (line 19–20). The UCB of item $i$ is estimated as

$$U_{t}(a) = \bar{\alpha}_{t}(\tau, a) + c_{t}(\tau, a),$$

where $\bar{\alpha}_{t}(\tau, a) = \frac{X_{t}(a)}{N_{t}(a)}$ is the average of observed attraction indicators of item $a$ inside the sliding window and

$$c_{t}(\tau, a) = \sqrt{\frac{\ln (t \wedge \tau)}{N_{t}(a)}}$$

is the confidence interval, and $t \wedge \tau = \min(t, \tau)$.

### Initialization
In the initialization phase, we set all the statistics to 0 and define $\tau := 1$ for any $x$ (line 3–4). Mapping back this to UCB, at the beginning, each item has the optimal assumption on the attraction probability with an optimal bonus on uncertainty. This is a common initialization strategy for UCB-type bandit algorithms [Zoghi et al., 2014; Li et al., 2018].

### 4 Analysis
In this section, we analyze the $n$-step regret of CascadeDUCB and CascadeSWUCB. We first derive regret upper bounds on CascadeDUCB and CascadeSWUCB, respectively. Then we derive a regret lower bound on cascading non-stationary bandits. Finally, we discuss our theoretical results.
4.1 Regret upper bound

We refer to $D_t^* \subseteq [L]$ as the set of the $K$ most attractive items in set $\mathcal{D}$ at time $t$ and $\mathcal{D}_t$ as the complement of $D_t^*$, i.e., $\forall a \in D_t^*, \forall a^* \in \mathcal{D}_t : \alpha_t(a) \geq \alpha_t(a^*)$ and $D_t^* \cup \mathcal{D}_t = \mathcal{D} ; D_t^* \cap \mathcal{D}_t = \emptyset$. At time $t$, we say an item $a^*$ is optimal if $a^* \in D_t^*$ and an item $a$ is suboptimal if $a \in \mathcal{D}_t$. The regret at time $t$ is caused by the case that $R_t$ includes at least one suboptimal and examined items. Let $\Delta_{a^*,a}$ be the gap of attraction probability between a suboptimal item $a$ and an optimal $a^*$ at time $t$: $\Delta_{a^*,a} = \alpha_t(a^*) - \alpha_t(a)$. Then we refer to $\Delta_{a,K}$ as the smallest gap of between item $a$ and the $K$-th most attractive item in all steps when $a$ is not the optimal item: $\Delta_{a,K} = \min_{t \in [n], a^* \in D_t^*} \alpha_t(a^*) - \alpha_t(a)$.

**Theorem 1.** Let $\epsilon \in (1/2, 1)$ and $\gamma \in (1/2, 1)$, the expected $n$-step regret of CascadeDUCB is bounded as:

$$ R(n) \leq L\gamma_n \frac{\ln(1-\gamma)\epsilon}{\ln \gamma} + \sum_{a \in \mathcal{D}} C(\gamma, a) \binom{n(1-\gamma)}{\ln \frac{1}{1-\gamma}} \ln \frac{1}{\gamma 1-\gamma}, \quad (9) $$

where

$$ C(\gamma, a) = \frac{4}{1-1/\epsilon} \ln(1 + 4 \sqrt{1 - 1/2\epsilon}) + \frac{32\epsilon}{\Delta_{a,K} \gamma(1-\gamma)} \cdot \quad (10) $$

We outline the proof in 4 steps below and the full version is in Appendix A.1.

**Proof.** Our proof is adapted from the analysis in [Kveton et al., 2015]. The novelty of the proof comes from the fact that, in a non-stationary environment, the discounted estimator $\alpha_t(\gamma, a)$ is now a biased estimator of $\alpha_t(a)$ (Step 1, 2 and 4).

Step 1. We bound the regret of the event that estimators of the attraction probabilities are biased by $L\gamma_n T_f$.

Step 2. We bound the regret of the event that $\alpha_t(a)$ falls outside of the confidence interval around $\alpha_t(\tau, a)$ by

$$ \ln^2 \tau + 2n \left[ \frac{\ln \tau}{\ln(1 + 4 \sqrt{1 - 1/2\epsilon})} \right]. \quad (14) $$

Step 3. We decompose the regret at time $t$ based on [Kveton et al., 2015, Theorem 1].

Step 4. For each item $a$, we bound the number of times that item $a$ is chosen when $a \in \mathcal{D}_t$ in $n$ steps and get the term $\frac{32\epsilon}{\Delta_{a,K} \gamma(1-\gamma)} \frac{\ln \frac{1}{\gamma}}{\ln \frac{1}{1-\gamma}}$. Finally, we sum up all the regret. \qed

The bound depends on $n$ steps and the number of breakpoints $T_f$. If they are known beforehand, we can choose $\gamma$ by minimizing the right hand side of Eq. 9. Choosing $\gamma = 1 - 1/4\sqrt{T_f/n}$ leads to $R(n) = O(\sqrt{nT_f \ln n})$. When $T_f$ is independent of $n$, we have $R(n) = O(\sqrt{nT_f \ln n})$.

**Theorem 2.** Let $\epsilon \in (1/2, 1)$. For any integer $\tau$, the expected $n$-step regret of CascadeSWUCB is bounded as:

$$ R(n) \leq L\gamma_n T_f + \frac{L\ln^2 \tau}{\ln(1 + 4 \sqrt{1 - 1/2\epsilon})} + \sum_{a \in \mathcal{D}} C(\tau, a) \frac{n\ln \tau}{\tau}, \quad (11) $$

where

$$ C(\tau, a) = \frac{2}{\ln \tau} \left[ \frac{\ln \tau}{\ln(1 + 4 \sqrt{1 - 1/2\epsilon})} \right] + \frac{8\epsilon}{\Delta_{a,K} \frac{n}{\tau}} \cdot \quad (12) $$

When $\tau$ goes to infinity and $n/\tau$ goes to 0,

$$ C(\tau, a) = \frac{2}{\ln(1 + 4 \sqrt{1 - 1/2\epsilon})} + \frac{8\epsilon}{\Delta_{a,K}}. \quad (13) $$

We outline the proof in 4 steps below and the full version is in Appendix A.2.

**Proof.** The proof follows the same lines as the proof of Theorem 1.

Step 1. We bound the regret of the event that estimators of the attraction probabilities are biased by $L\gamma_n T_f$.

Step 2. We bound the regret of the event that $\alpha_t(a)$ falls outside of the confidence interval around $\alpha_t(\tau, a)$ by

$$ \ln^2 \tau + 2n \left[ \frac{\ln \tau}{\ln(1 + 4 \sqrt{1 - 1/2\epsilon})} \right]. \quad (14) $$

Step 3. We decompose the regret at time $t$ based on [Kveton et al., 2015, Theorem 1].

Step 4. For each item $a$, we bound the number of times that item $a$ is chosen when $a \in \mathcal{D}_t$ in $n$ steps and get the term $\frac{32\epsilon}{\Delta_{a,K} \gamma(1-\gamma)} \frac{\ln \frac{1}{\gamma}}{\ln \frac{1}{1-\gamma}}$. Finally, we sum up all the regret. \qed

If we know $T_f$ and $n$ beforehand, we can choose the window size $\tau$ by minimizing the right hand side of Eq. 11. Choosing $\tau = 2\sqrt{n \ln(n)/T_f}$ leads to $R(n) = O(\sqrt{n T_f \ln n})$. When $T_f$ is independent of $n$, we have $R(n) = O(\sqrt{n T_f \ln n})$.

4.2 Regret lower bound cascading non-stationary bandits

We consider a particular cascading non-stationary bandit and refer to it as $B_L = (L, K, \Delta, p, \Upsilon)$. We have a set of $L$ items $\mathcal{D} = [L]$ and $K = \frac{1}{2}L$ positions. At any time $t$, the distribution of attraction probability of each item $a \in \mathcal{D}$ is parameterized by:

$$ \alpha_t(a) = \begin{cases} p & \text{if } a \in D_t^* \\ p - \Delta & \text{if } a \in \mathcal{D}_t, \end{cases} \quad (15) $$

where $D_t^*$ is the set of optimal items at time $t$, $\mathcal{D}_t$ is the set of suboptimal items at time $t$, and $\Delta \in (0, p)$ is the gap between optimal items and suboptimal items. Thus, the attraction probabilities only take two values: $p$ for optimal items and $p - \Delta$ for suboptimal items up to $n$-step. $\Upsilon$ is the number of breakpoints when the attraction probability of an item changes from $p$ to $p - \Delta$ or other way around. Particularly, we consider a simple variant that the distribution of attraction probability of each item is piecewise constant and has two breakpoints. And we assume another constraint on the number of optimal items that $|D_t^*| = K$ for all time steps $t \in [n]$. Then, the regret that any learning policy can achieve when interacting with $B_L$ is lower bounded by Theorem 3.
Theorem 3. The n-step regret of any learning algorithm interacting with cascading non-stationary bandit $B_L$ is lower bounded as follows:

$$\liminf_{n \to \infty} R(n) \geq L \Delta(1-p)^{K-1} \sqrt{\frac{2n}{3D_{KL}(p - \Delta||p)}}, \quad \text{(16)}$$

where $D_{KL}(p - \Delta||p)$ is the Kullback-Leibler (KL) divergence between two Bernoulli distributions with means $p - \Delta$ and $p$.

Proof. The proof is based on the analysis in [Kveton et al., 2015]. We first refer to $R_t^*$ as the optimal list at time $t$ that includes $K$ items. For any time step $t$, any item $a \in D_t$ and any item $a^* \in D_t^*$, we define the event that item $a$ is included in $R_t$ instead of item $a^*$ and item $a$ is examined but not clicked at time step $t$ by:

$$G_{t,a,a^*} = \{1 \leq k \leq c_t \ s.t. \ R_t(k) = a, R_t(k) = a^*\}. \quad \text{(17)}$$

By [Kveton et al., 2015, Theorem 1], the regret at time $t$ is decomposed as:

$$\mathbb{E}[r(R_t, \alpha_t)] \geq \Delta(1-p){K-1} \sum_{a \in D_t} \sum_{a^* \in D_t^*} 1\{G_{t,a,a^*}\}. \quad \text{(18)}$$

Then, we bound the $n$-step regret as follows:

$$R(n) \geq \Delta(1-p){K-1} \sum_{t=1}^{n} \sum_{a \in D_t} 1\{G_{t,a,a^*}\} \geq \Delta(1-p){K-1} \sum_{t=1}^{n} \sum_{a \in D_t} 1\{a \in D_t, a \in R_t\}$$

$$= \Delta(1-p){K-1} \sum_{a \in D} T_n(a), \quad \text{(19)}$$

where $T_n(a) = \sum_{t=1}^{n} 1\{a \in D_t, a \in R_t, R_{t-1}(a) \leq c_t\}$. The second inequality is based on the fact that, at time $t$, the event $G_{t,a,a^*}$ happens if and only if item $a$ is suboptimal and examined. By the results of [Garivier and Moulines, 2011, Theorem 3], if a suboptimal item $a$ has not been examined enough times, the learning policy may play this item for a long period after a breakpoint. And we get:

$$\liminf_{n \to \infty} T(n) \geq \sqrt{\frac{2n}{3D_{KL}(p - \Delta||p)}}. \quad \text{(20)}$$

We conclude the proof by summing up all the inequalities and get:

$$\liminf_{n \to \infty} R(n) \geq L \Delta(1-p)^{K-1} \sqrt{\frac{2n}{3D_{KL}(p - \Delta||p)}}. \quad \square$$

4.3 Discussion

We have shown that the $n$-step regret upper bounds of CascadeDUCB and CascadeSWUCB have the order of $O(\sqrt{n} \ln n)$ and $O(\sqrt{n} \ln \ln n)$, respectively. They all match the lower bound proposed in Theorem 3 up to a logarithm factor. Specifically, the upper bound of CascadeDUCB matches the lower bound up to $\ln n$. The upper bound of CascadeSWUCB matches the lower bound up to $\sqrt{\ln n}$, which is slightly better than CascadeDUCB and has been confirmed by experiments in Section 5.

The above discussion is under the assumption that step $n$ is known beforehand. Practically, this is not always possible. We can extend CascadeDUCB and CascadeSWUCB to the case where $n$ is unknown by using the doubling trick [Garivier and Moulines, 2011]. Namely, for $t > n$ and any $k$, such that $2^k \leq t < 2^{k+1}$, we reset $c_t = \frac{1}{\sqrt{2^{k+1}}}$ and $\tau = 2\sqrt{\ln(2^k)}$.

Both CascadeDUCB and CascadeSWUCB can be computed efficiently. Their computational complexity is linear in the number of time steps. However, CascadeSWUCB requires extra memory to remember the past $\tau$ ranked lists and rewards for updating $X_t$ and $N_t$.}

5 Experimental Analysis

We evaluate CascadeDUCB and CascadeSWUCB on the Yandex click dataset, which is the largest public click collection. It contains more than 30 million search sessions, each of which contains at least one search query. We process the queries in the same manner as in [Zoghi et al., 2017; Li et al., 2019]. Namely, we randomly select 100 frequent search queries with the 10 most attractive items in each query, and then learn a CM for each query using PyClick. These CMs are used to generate click feedback. In this setup, the size of candidate items is $L = 10$ and we choose $K = 3$ as the number of positions. The objective of the learning task is to identify 3 most attractive items and then maximize the expected number of clicks at the 3 highest positions.

We consider a simulated non-stationary environment setup, where we take the learned attraction probabilities as the default and change the attraction probabilities periodically. Our simulation can be described in 4 steps: (1) For each query, the attraction probabilities of the top 3 items remain constant over time. (2) We randomly choose three suboptimal items and set their attraction probabilities to 0.9 for the next $m_1$ steps. (3) Then we reset the attraction probabilities and keep them constant for the next $m_2$ steps. (4) Repeat step (2) and step (3) iteratively. This simulation mimics abrupt changes in user preferences and is widely used in previous work on non-stationarity [Garivier and Moulines, 2011; Wu et al., 2018; Jagerman et al., 2019]. In our experiment, we set $m_1 = m_2 = 10K$ and choose 10 breakpoints. In total, we run experiments for 100K steps. Although the non-stationary aspects in our setup are simulated, the other parameters of a CM are learned from the Yandex click dataset.

We compare CascadeDUCB and CascadeSWUCB to RankedEXP3 [Radlinski et al., 2008], CascadeKL-UCB [Kveton et al., 2015] and BatchRank [Zoghi et al., 2017]. We describe the baseline algorithms in slightly more details in Section 6. Briefly, RankedEXP3, a variant of ranked bandits, is based on an adversarial bandit algorithm Exp3 [Auer et al., 1995]; it is the earliest bandit-based ranking algorithm and is popular in practice. CascadeKL-UCB [Kveton et al., 2015]
is a near optimal algorithm in CM. BatchRank [Zoghi et al., 2017] can learn in a wide range of click models. However, these algorithms only learn in a stationary environment. We choose them as baselines to show the superiority of our algorithms in a non-stationary environment. In experiments, we set ϵ = 0.5, γ = 1 − 1/(4√n) and τ = 2√n ln(n), the values that roughly minimize the upper bounds.

We report the n-step regret averaged over 100 queries and 10 runs per query in Figure 1. We see that all baselines have linear regret in time step. They fail in capturing the breakpoints. What is even worse is that the non-stationarity makes the baselines perform even worse during epochs where the attraction probability are set as the default. For example, CascadeKL-UCB has 111.50 ± 1.12 regret in the first 10k steps but has 447.82 ± 137.16 regret between step 80k and 90k. We emphasize that the attraction probabilities equal the default and remain constant inside these two epochs. This result is caused by the fact that the baseline algorithms rank items based on all historical observations, i.e., they do not balance “remembering” and “forgetting.” Because of the use of a discounting factor and/or a sliding window, CascadeDUCB and CascadeSWUCB can detect breakpoints and show convergence. CascadeSWUCB outperforms CascadeDUCB with a small gap. This observation is consistent with our theoretical finding that CascadeSWUCB outperforms CascadeDUCB by a \( \sqrt{\ln n} \) factor.

### 6 Related Work

The idea of directly learning to rank from user feedback has been widely studied in a stationary environment. Ranked bandits [Radlinski et al., 2008; Slivkins et al., 2013] are among the earliest OLTR approaches. In ranked bandits, each position in the list is modeled as an individual underlying MABs. The ranking task is then solved by asking each individual MAB to recommend an item to the attached position. Since the reward, e.g., click, of a lower position is affected by higher positions, the underlying MAB is typically adversarial, e.g., Exp3 [Auer et al., 1995; Auer et al., 2002b]. BatchRank is a recently proposed OLTR method [Zoghi et al., 2017]; it is an elimination-based algorithm: once an item is found to be inferior to \( K \) items, it will be removed from future consideration. BatchRank outperforms ranked bandits in the stationary environment. In our experiments, we include BatchRank and RankedEXP3, the Exp3-based ranked bandit algorithm, as baselines.

Several OLTR algorithms have been proposed in specific click models [Kveton et al., 2015; Lagrée et al., 2016; Katariya et al., 2016; Oosterhuis and de Rijke, 2018]. They can efficiently learn the optimal ranking given the click model they consider. Our work is related to cascading bandits and we compare our algorithms to CascadeKL-UCB, an algorithm proposed for solving cascading bandits [Kveton et al., 2015], in Section 5. Our work differs from cascading bandits in that we consider learning in a non-stationary environment.

Non-stationary bandit problems have been widely studied in the literature [Slivkins and Upfal, 2008; Yu and Mannor, 2009; Gabrie and Mounilines, 2011; Besbes et al., 2014; Liu et al., 2018]. However, previous work requires a small action space. In our OLTR setup, the actions are ranked lists, of which the number is exponential in \( K \). Thus, we do not consider them as baselines in our experiments.

Another closely related topic in MABs is adversarial bandits, where the reward realizations, in our case attraction indicators, at any time step are selected by an adversary. Adversarial bandits originate from game theory [Blackwell, 1956] and have been widely studied, cf. [Auer et al., 1995; Cesa-Bianchi and Lugosi, 2006] for an overview. Within adversarial bandits, the performance of a policy is often measured by comparing to a static oracle which always chooses a single best arm that is obtained after seeing all the reward realizations up to step \( n \). This static oracle can perform poorly in a non-stationary case when the single best arm is suboptimal for a long time between two breakpoints. Thus, even if a policy performs closely to the static oracle, it can still perform sub-optimally in a non-stationary environment. Our work differs from adversarial bandits in that we compare to a dynamic oracle that can balance the dilemma of “remembering” and “forgetting” and chooses the per-step best action.

### 7 Conclusion

In this paper, we have studied the online learning to rank (OLTR) problem in a non-stationary environment where user preferences change abruptly. We focus on a widely-used user click behavior model cascade model (CM) and have proposed an online learning variant of it called cascading non-stationary bandits. Two algorithms, CascadeDUCB and CascadeSWUCB, have been proposed for solving it. Our theoretical have shown that they have sub-linear regret. These theoretical findings have been confirmed by our experiments on the Yandex click dataset. We open several future directions for non-stationary OLTR. First, we have only considered the CM setup. Other click models that can handle multiple clicks, e.g., DBN [Chapelle and Zhang, 2009], should be considered as part of future work. Second, we focused on UCB-based policy. Another possibility is to use the family of softmax policies [Besbes et al., 2014]. Along this line, one may obtain upper bounds independent from the number of breakpoints.
References


A Proofs

In the appendix, we refer to $\mathcal{R}^*_t$ as the optimal list at time $t$ that includes $K$ items sorted by the decreasing order of their attraction probabilities. We refer to $\mathcal{D}^*_t \subseteq [L]$ as the set of the $K$ most attractive items in set $\mathcal{D}$ at time $t$ and $\mathcal{D}_t$ as the complement of $\mathcal{D}^*_t$, i.e., $\forall a^* \in \mathcal{D}^*_t, \forall a \in \mathcal{D}_t : \alpha_t(a^*) \geq \alpha_t(a)$ and $\mathcal{D}^*_t \cup \mathcal{D}_t = \mathcal{D}, \mathcal{D}^*_t \cap \mathcal{D}_t = \emptyset$. At time $t$, we say an item $a^*$ is optimal if $a^* \in \mathcal{D}^*_t$ and an item $a$ is suboptimal if $a \in \mathcal{D}_t$. We denote $r_t = \max_{R \in \Pi_K(\mathcal{D})} r(R, \alpha_t) - r(R_t, A_t)$ to be the regret at time $t$ of the learning algorithm. Let $\Delta'_{a,a^*}$ be the gap of attraction probability between a suboptimal item $a$ and an optimal item $a^*$ at time $t$: $\Delta'_{a,a^*} = \alpha_t(a^*) - \alpha_t(a)$. Then we refer to $\Delta'_{a,K}$ as the smallest gap between item $a$ and the $K$-th most attractive item in all $n$ time steps when $a$ is not the optimal item: $\Delta_{a,K} = \min_{a \in [n], a \notin \mathcal{D}^*_t} \alpha_t(K) - \alpha_t(a)$.

A.1 Proof of Theorem 1

Let $M_t = \{ \exists a \in \mathcal{D} \text{ s.t. } |\alpha_t(a) - \bar{\alpha}_t(\gamma, a)| > c_t(\gamma, t) \}$ be the event that $\alpha_t(a)$ falls out of the confidence interval around $\bar{\alpha}_t(\gamma, a)$ at time $t$, and $M_t$ be the complement of $M_t$. We re-write the n-step regret of CascadeDUCB as follows:

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} \mathbbm{1}\{M_t\} r_t \right] + \mathbb{E} \left[ \sum_{t=1}^{n} \mathbbm{1}\{\bar{M}_t\} r_t \right].$$

We then bound both terms above separated.

We refer to $\mathcal{T}$ as the set of all time steps such that for $t \in \mathcal{T}, s = [t - B(\gamma), t]$ and any item $a \in \mathcal{D}$ we have $\alpha(s(a)) = \alpha_t(a)$, where $B(\gamma) = \lceil \log_2(\epsilon + 1 - \gamma) \rceil$. In other words, $\mathcal{T}$ is the set of time steps that do not follow too close after breakpoints. Since for any time step $t \notin \mathcal{T}$ the estimators of attraction probabilities are biased, CascadeDUCB may recommend suboptimal items constantly. Thus, we get the following bound:

$$\mathbb{E} \left[ \sum_{t=1}^{n} \mathbbm{1}\{M_t\} r_t \right] \leq L Y_n B(\gamma) + \mathbb{E} \left[ \sum_{t \in \mathcal{T}} \mathbbm{1}\{M_t\} r_t \right].$$

Then, we show that for steps $t \in \mathcal{T}$, the attraction probabilities are well estimated for all items with high probability. For an item $a$, we consider the bias and variance of $\bar{x}_t(\gamma, a)$ separately. We denote:

$$X_t(\gamma, a) = \sum_{s=1}^{t} x(s, \alpha_t(a) - \bar{x}_t(\gamma, a)) \mathbbm{1}\{a \in R_s, R_s(c_s) = a\}, \quad N_t(\gamma, a) = \sum_{s=1}^{t} x(s, \alpha_t(a) - \bar{x}_t(\gamma, a)) \mathbbm{1}\{a \in R_s\},$$

as the discounted number of being clicked, and the discounted number of being examined.

First, we consider the bias. The “bias” at time $t$ can be written as $x_t(\gamma, a)/N_t(\gamma, a) - \alpha_t(a)$, where $x_t(\gamma, a) = \sum_{s=1}^{t} x(s, \alpha_t(a) - \bar{x}_t(\gamma, a)) \mathbbm{1}\{a \in R_s\} \alpha_t(a)$. For $t \in \mathcal{T}$:

$$|x_t(\gamma, a) - \alpha_t(a)| = \left| \sum_{s=1}^{t-B(\gamma)} x(s, \alpha_t(a) - \bar{x}_t(\gamma, a)) \mathbbm{1}\{a \in R_s\} \right| \leq \left| \sum_{s=1}^{t-B(\gamma)} x(s, (\alpha_t(a) - \bar{x}_t(\gamma, a)) \mathbbm{1}\{a \in R_s\} \right| \leq \gamma B(\gamma) N_{t-B(\gamma)}(\gamma, a) \leq \frac{\gamma B(\gamma)}{1 - \gamma},$$

where the last inequality is due to the fact that $N_{t-B(\gamma)}(\gamma, a) \leq 1/(1 - \gamma)$. Thus, we get:

$$\left| \frac{x_t(\gamma, a)}{N_t(\gamma, a)} - \alpha_t(a) \right| \leq \frac{\gamma B(\gamma)}{(1 - \gamma) N_t(\gamma, a)} \leq \left( \frac{\gamma B(\gamma)}{1 - \gamma} \right) \frac{\ln N_t(\gamma, a)}{N_t(\gamma, a)} \leq \frac{1}{2} c_t(\gamma, a),$$

where the third inequality is due to the fact that $\ln x \leq \sqrt{x}$ and the last inequality is due to the definition of $B(\gamma)$. So, $B(\gamma)$ time steps after a breakpoint, the “bias” is at most half as large as the confidence bonus; and the second half of the confidence interval is used for the variance.

Second, we consider the variance. For $a \in \mathcal{D}$ and $t \in \mathcal{T}$, let $M_{t,a} = \{ |\alpha_t(a) - \bar{x}_t(\gamma, a)| > c_t(\gamma, t) \}$ be the event that $\alpha_t(a)$ falls out of the confidence interval around $\bar{x}_t(\gamma, a)$ at time $t$. By using a Hoeffding-type inequality (Garivier and Moulines,
2011, Theroem 4]), for an item \( a \in D \), \( t \in T \), and any \( \eta > 0 \), we get:

\[
P(M_{t,a}) \leq P \left( \frac{X_t(\gamma, a) - x_t(\gamma, a)}{\sqrt{N_t(\gamma^2, a)}} > \sqrt{\frac{\epsilon \ln N_t(\gamma)}{N_t(\gamma^2, a)}} \right)
\]

\[
\leq P \left( \frac{X_t(\gamma, a) - x_t(\gamma, a)}{\sqrt{N_t(\gamma^2, a)}} > \sqrt{\epsilon \ln N_t(\gamma)} \right)
\]

\[
\leq \left[ \frac{\ln N_t(\gamma)}{\ln (1 + \eta)} \right] N_t(\gamma)^{-2\epsilon(1-\frac{\eta^2}{\bar{a}^2})}.
\]

Thus, we get the following bound:

\[
\mathbb{E} \left[ \sum_{t \in T} 1\{M_t\} r_t \right] \leq 2L \sum_{t \in T} \left[ \frac{\ln N_t(\gamma)}{\ln (1 + \eta)} \right] N_t(\gamma)^{-2\epsilon(1-\frac{\eta^2}{\bar{a}^2})}.
\]

By taking \( \eta = 4\sqrt{1-2/\epsilon} \) such that \( 1 - \frac{\eta^2}{\bar{a}^2} = 1 \), and with \( t_0 = (1 - \gamma)^{-1} \) we get:

\[
\sum_{t \in T} \left[ \frac{\ln N_t(\gamma)}{\ln (1 + \eta)} \right] N_t(\gamma)^{-2\epsilon(1-\frac{\eta^2}{\bar{a}^2})} \leq t_0 + \sum_{t \in T, t \geq t_0} \left[ \frac{\ln N_{t_0}(\gamma)}{\ln (1 + \eta)} \right] N_{t_0}(\gamma)^{-1}
\]

\[
\leq t_0 + \left[ \frac{\ln N_{t_0}(\gamma)}{\ln (1 + \eta)} \right] \frac{n}{N_{t_0}(\gamma)}
\]

\[
\leq \frac{1}{1 - \gamma} + \left[ \frac{\ln N_{t_0}(\gamma)}{\ln (1 + \eta)} \right] \frac{n(1 - \gamma)}{1 - \gamma^{1/(1-\gamma)}}.
\]

We sum up and get the upper bound:

\[
\mathbb{E} \left[ \sum_{i=1}^n 1\{M_i\} r_i \right] \leq L \Upsilon_n B(\gamma) + 2L \frac{1}{1 - \gamma} + 2L \left[ \frac{\ln N_{t_0}(\gamma)}{\ln (1 + \eta)} \right] \frac{n(1 - \gamma)}{1 - \gamma^{1/(1-\gamma)}}.
\]

(23)

Third, we upper bound the second term in Eq. 21. The regret is caused by recommending a suboptimal item to the user and the user examines but does not click the item. Since there are \( \Upsilon_n \) breakpoints, we refer to \( [t_1, \ldots, t_{\Upsilon_n}] \) as the time step of a breakpoint that occurs. We consider the time step in the individual epoch that does not contain a breakpoint. For any epoch and any time \( t \in \{t_\epsilon, t_\epsilon + 1, \ldots, t_{\epsilon+1} - 1\} \), any item \( a \in \bar{D}_\epsilon \) and any item \( a^* \in D^*_\epsilon \), we define the event that item \( a \) is included in \( D_t \) instead of item \( a^* \), and item \( a \) is examined but not clicked at time \( t \) by:

\[
G_{t,a,a^*} = \{ \exists 1 \leq k < c_t \ s.t. \ R_t(k) = a, R_t(k) = a^* \}.
\]

Since the attraction probability remains constant in the epoch, we refer to \( D^*_\epsilon \) as the optimal items and \( \bar{D}_\epsilon \) as the suboptimal items in epoch \( e \). By [Kveton et al., 2015, Theorem 1], the regret at time \( t \) is decomposed as:

\[
\mathbb{E}[r_t] \leq \Delta_t^{a^*,a} \sum_{a \in \bar{D}_\epsilon} \sum_{a^* \in D^*_\epsilon} 1\{G_{a,a^*,t}\}.
\]

(24)

Then we have:

\[
\mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1}-1} 1\{M_t\} r_t \right] = \sum_{t=t_i}^{t_{i+1}-1} \mathbb{E} \left[ 1\{M_t\} r_t \right] \leq \sum_{a \in \bar{D}_\epsilon} \sum_{a^* \in D^*_\epsilon} \mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1}-1} \Delta_t^{a,a^*} 1\{G_{a,a^*,t}\} \right],
\]

(25)

where the first equality is due to the tower rule, and the inequality is due to Eq. 24.

Now, for any suboptimal item \( a \) in epoch \( e \), we upper bound \( \sum_{a^* \in D^*_\epsilon} \sum_{t=t_i}^{t_{i+1}-1} \Delta_t^{a,a^*} 1\{G_{a,a^*,t}\} \). At time \( t \), event \( 1\{M_t\} \) and event \( 1\{a \in R_t, a \notin \bar{D}_t\} \) happen when there exists an optimal item \( a^* \in D^*_\epsilon \) such that:

\[
\alpha_t(a) + 2\epsilon_t(\gamma, a) \geq U_t(a) \geq U(a^*) \geq \alpha_t(a^*),
\]

which implies that \( 2\epsilon_t(\gamma, a) \geq \alpha_t(a^*) - \alpha_t(a) \). Taking the definition of the confidence interval, we get:

\[
N_t(\gamma, a) \leq \frac{16\epsilon \ln N_t(\gamma)}{\Delta_t^{a,a^*}},
\]
where we set $\Delta_{t,a,a^*} = \Delta_{t,a^*}$. Together with Eq. 25, we get:

\[
E \left[ \sum_{t=t_i}^{t_i+1} 1 \{ \bar{M}_t \} r_t \right] \leq \sum_{a \in D} E \left[ \sum_{a^* \in D^*} \frac{16\epsilon \ln N_t(\gamma)}{\Delta_{t,a,a^*}} \right]
\]

\[
\leq 16\epsilon \ln N_t(\gamma) \left[ \Delta_{t,a,1} \frac{1}{\Delta_{t,a,1}} + \sum_{a^* = 2}^{K} \Delta_{t,a,a^*} \left( \frac{1}{\Delta_{t,a,a^*}} - \frac{1}{\Delta_{t,a,a^*} - 1} \right) \right]
\]

\[
\leq \frac{32\epsilon \ln N_t(\gamma)}{\Delta_{t,a,K}},
\]

(26)

where the last inequality is due to [Kveton et al., 2014, Lemma 3]. Let $\Delta_{a,K} = \min_{t \in [n]} \Delta_{t,a,K}$ be the smallest gap between the suboptimal item $a$ and an optimal item in all time steps. When $N_t(\gamma, a) > \frac{32\epsilon \ln N_t(\gamma)}{\Delta_{a,K}}$, CascadeDUCB will not select item $a$ at time $t$. Thus we get:

\[
\sum_{a \in D} E \left[ \sum_{t=1}^{n} 1 \{ \bar{M}_t \} 1 \{ a \in R_t, a \in D_t \} \right] \leq \sum_{a \in [\Upsilon_t]} \sum_{a \in D} \frac{32\epsilon \ln N_t(\gamma)}{\Delta_{t,a,K}}
\]

\[
\leq \sum_{a \in D} \left[ n(1 - \gamma) \right] \frac{32\epsilon \ln N_n(\gamma)}{\Delta_{a,K}} \gamma^{1/(1 - \gamma)},
\]

(27)

where the last inequality is based on [Garivier and Moulines, 2011, Lemma 1]. Finally, together with Eq. 21, Eq. 22, Eq. 23, Eq. 24, Eq. 25 and Eq. 27, we get Theorem 1.

A.2 Proof of Theorem 2

Let $M_t = \{ \exists a \in D \text{ s.t. } |\alpha_t(a) - \bar{\alpha}_t(\tau, a)| > c_t(\tau, t) \}$ be the event that $\alpha_t(a)$ falls out of the confidence interval around $\bar{\alpha}_t(\tau, a)$ at time $t$, and let $\bar{M}_t$ be the complement of $M_t$. We re-write the $n$-step regret of CascadeSWUCB as follows:

\[
R(n) = E \left[ \sum_{t=1}^{n} 1 \{ M_t \} r_t \right] + E \left[ \sum_{t=1}^{n} 1 \{ \bar{M}_t \} r_t \right].
\]

(28)

We then bound both terms in Eq. 28 separately. First, we refer to $\mathcal{T}$ as the set of all time steps such that for $t \in \mathcal{T}, s \in [t - \tau, t]$ and any item $a \in D$ we have $\alpha_s(a) = \alpha_t(a)$. In other words, $\mathcal{T}$ is the set of time steps that do not follow too close after breakpoints. Obviously, for any time step $t \notin \mathcal{T}$ the estimators of attraction probabilities are biased and CascadeSWUCB may recommend suboptimal items constantly. Thus, we get the following bound:

\[
E \left[ \sum_{t=1}^{n} 1 \{ M_t \} r_t \right] \leq L\mathcal{Y_n} + E \left[ \sum_{t \in \mathcal{T}} 1 \{ M_t \} r_t \right].
\]

(29)

$\tau$ time steps after a breakpoint, the estimators of the attraction probabilities are not biased.

Then, we consider the variance. By using a Hoeffding-type inequality ([Garivier and Moulines, 2008, Corollary 21]), for an item $a \in D, t \in \mathcal{T}$, and any $\eta > 0$, we get:

\[
P \left( |\hat{\alpha}_t(\tau, a) - \alpha_t(a)| > c_t(\tau,t) \right) \leq P \left( \hat{\alpha}_t(\tau, a) > \alpha_t(a) + \frac{\epsilon \ln (t \wedge \tau)}{N_t(\tau, a)} \right) + P \left( \hat{\alpha}_t(\tau, a) < \alpha_t(a) - \frac{\epsilon \ln (t \wedge \tau)}{N_t(\tau, a)} \right)
\]

\[
\leq 2 \left[ \frac{\ln (t \wedge \tau)}{\ln(1 + \eta)} \right] \exp \left( -2\epsilon \ln (t \wedge \tau)(1 - \frac{\eta}{16}) \right)
\]

\[
= 2 \left[ \frac{\ln (t \wedge \tau)}{\ln(1 + \eta)} \right] (t \wedge \tau)^{-2\epsilon(1 - \eta^2/16)}.
\]

Taking $\eta = 4\sqrt{1 - \frac{1}{2\epsilon}}$, we have: $P \left( |\hat{\alpha}_t(\tau, a) - \alpha_t(a)| \right) \leq 2 \left[ \frac{\ln (t \wedge \tau)}{t \wedge \tau} \right].$ Thus, we get the following bound:

\[
E \left[ \sum_{t \in \mathcal{T}} 1 \{ M_t \} r_t \right] \leq 2L \sum_{t \in \mathcal{T}} \left[ \frac{\ln (t \wedge \tau)}{t \wedge \tau} \right] \leq \frac{L \ln^2(\tau)}{\ln(1 + 4\sqrt{1 - 1/2\epsilon})} + \frac{2L \ln(\tau)}{\tau \ln(1 + 4\sqrt{1 - 1/2\epsilon})}.
\]
We sum up and get the upper bound:

$$
\mathbb{E} \left[ \sum_{t=1}^{n} \mathbb{I} \{ M_t \} R_t \right] \leq L Y_n^2 + \frac{L \ln^2(\tau)}{\ln(1 + 4\sqrt{1 - 1/2\epsilon})} + \frac{2L \ln \tau}{\tau \ln(1 + 4\sqrt{1 - 1/2\epsilon})}.
$$

(30)

Third, we upper bound the second term in Eq. 28. The regret is caused by recommending a suboptimal item to the user and the user examines but does not click the item. Since there are $Y_n$ breakpoints, we refer to $[t_1, \ldots, t_{Y_n}]$ as the time steps of a breakpoint. We consider the time step in the individual epoch that does not contain a breakpoint. For any epoch and any time $t \in \{ t_e, t_e + 1, \ldots, t_{e+1} - 1 \}$, any item $a \in \mathcal{D}_e$ and any item $a^* \in \mathcal{D}_e^*$, we define the event that item $a$ is included in $\mathbf{R}_t$ instead of item $a^*$ and item $a$ is examined but not clicked at time $t$ by:

$$
G_{t,a,a^*} = \{ \exists 1 \leq k < c_t \text{ s.t. } \mathbf{R}_t(k) = a, \mathbf{R}_t(k) = a^* \}.
$$

Since the attraction probabilities remain constant in the epoch, we refer to $\mathcal{D}_e^*$ as the optimal items and $\mathcal{D}_e$ as the suboptimal items in epoch $e$. By [Kveton et al., 2015, Theorem 1], the regret at time $t$ is decomposed as:

$$
\mathbb{E}[r_t] \leq \Delta_{t,a,a^*} \sum_{a \in \mathcal{D}_e} \sum_{a^* \in \mathcal{D}_e^*} 1 \{ G_{a,a^*,t} \},
$$

(31)

Then we have:

$$
\mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1} - 1} \mathbb{I} \{ M_t \} R_t \right] = \sum_{t=t_i}^{t_{i+1} - 1} \mathbb{E} \left[ \sum_{a \in \mathcal{D}_e} \sum_{a^* \in \mathcal{D}_e^*} \Delta_{t,a,a^*} \mathbb{I} \{ G_{a,a^*,t} \} \right],
$$

(32)

where the first equality is due to the tower rule, and the inequality if due to Eq. 31.

Now, for any suboptimal item $a$ in epoch $e$, we upper bound $\mathbb{E} \left[ \sum_{a^* \in \mathcal{D}_e^*} \sum_{t=t_i}^{t_{i+1} - 1} \mathbb{I} \{ G_{a,a^*,t} \} \right]$ at time $t$, event $\mathbb{I} \{ a \in \mathbf{R}_t, a \in \mathcal{D}_t \}$ happen when there exists an optimal item $a^* \in \mathcal{D}_e^*$ such that:

$$
\alpha_t(a) + 2\epsilon_t(\tau, a) \geq \mathbb{U}_t(a) \geq \mathbb{U}(a^*) \geq \alpha_t(a^*),
$$

which implies that $2\epsilon_t(\tau, a) \geq \alpha_t(a^*) - \alpha_t(a)$. Taking the definition of the confidence interval, we get:

$$
\mathbf{N}_t(\tau, a) \leq \frac{4\epsilon_t N_t(\tau)}{\Delta_{t,a,a^*}^2},
$$

where we set $\Delta_{t,a,a^*} = \Delta_{t,a,a^*}$. Together with Eq. 32, we get:

$$
\mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1} - 1} \mathbb{I} \{ M_t \} R_t \right] \leq \sum_{a \in \mathcal{D}_e} \mathbb{E} \left[ \sum_{a^* \in \mathcal{D}_e^*} 4\epsilon_t N_t(\gamma) \frac{1}{\Delta_{t,a,a^*}^2} \right]
\leq 4\epsilon_t N_t(\gamma) \left[ \Delta_{t,a,1} \frac{1}{\Delta_{t,a,1}^2} + \sum_{a^*=2}^{K} \Delta_{t,a,a^*} \left( \frac{1}{\Delta_{t,a,a^*}^2} - \frac{1}{\Delta_{t,a,a^*}^2 - 1} \right) \right]
\leq \frac{8\epsilon_t N_t(\gamma)}{\Delta_{t,a,K}},
$$

(33)

where the last inequality is due to [Kveton et al., 2014, Lemma 3]. Let $\Delta_{a,K} = \min_{t \in [n]} \Delta_{t,a,K}$ be the smallest gap between the suboptimal item $a$ and an optimal item in all time steps. When $\mathbf{N}_t(\tau, a) > \frac{8\epsilon_t N_t(\tau)}{\Delta_{a,K}}$, CascadeDUCB will not select item $a$ at time $t$. Thus we get:

$$
\sum_{a \in \mathcal{D}} \mathbb{E} \left[ \sum_{t=1}^{n} \mathbb{I} \{ M_t \} \mathbb{I} \{ a \in \mathbf{R}_t, a \in \mathcal{D}_t \} \right] \leq \sum_{a \in \mathbf{T}_n} \sum_{a \in \mathcal{D}_e} \frac{8\epsilon_t N_t(\tau)}{\Delta_{t,a,K}}
\leq \sum_{a \in \mathcal{D}} \left[ n \right] \frac{8\epsilon_t N_t(\tau)}{\Delta_{a,K}},
$$

(34)

where the last inequality is based on [Garivier and Moulines, 2008, Lemma 25].

Finally, together with Eq. 28, Eq. 29, Eq. 30, Eq. 32 and Eq. 34, we get Theorem 2.

**B Further Experiments**

In this section, we compare CascadeDUCB, CascadeSWUCB and baselines on single queries. We pick 20 queries and report the results in Figure 2. The results exemplify that CascadeDUCB and CascadeSWUCB have sub-linear regret while other baselines have linear regret.
Figure 2: The n-step regret of CascadeDUCB (black), CascadeSWUCB (red), RankedEXP3 (cyan), CascadeKL-UCB (green) and BubbleRank (blue) on single queries in up to 100k steps. Lower is better. The results are averaged over 10 runs per query. The shaded regions represent standard errors of our estimates.