On the Value of Prior in Online Learning to Rank

Branislav Kveton
Amazon*

Ofer Meshi
Google Research

Zhen Qin
Google Research

Masrour Zoghi
Google Research

Abstract

This paper addresses the cold-start problem in *online learning to rank (OLTR)*. We show both theoretically and empirically that priors improve the quality of ranked lists presented to users interactively based on user feedback. These priors can come in the form of unbiased estimates of the relevance of the ranked items, or more practically, can be obtained from offline-learned models. Our experiments show the effectiveness of priors in improving the short-term regret of tabular OLTR algorithms, based on Thompson sampling and BayesUCB.

1 INTRODUCTION

Learning to rank (LTR) is an important problem with many applications, such as in search (Liu, 2011), recommender systems (Falk, 2019), and ad placement (Tagami et al., 2013). The goal is to present a set of items to a population of users to optimize some utility, be it effort to find a piece of information, satisfaction with recommended results, or engagement with ads. Given the dynamic nature of user needs and desires, numerous *online learning to rank (OLTR)* algorithms have been proposed to learn from user feedback and adapt to user preferences (Radlinski et al., 2008; Jagerman et al., 2019; Wang et al., 2019).

As with most online learning methods, one major drawback of OLTR algorithms is the cold-start problem (Li et al., 2019). To guarantee asymptotic optimality, they need to explore at a certain rate. Although this rate typically tapers off as they see more data, they are generally overly exploratory earlier in the learning process. This early phase of aggressive exploration is often a hindrance to deployment of OLTR algorithms in practice. This is because a well-performing offline policy often exists, which is preferable to the exploration that OLTR algorithms inflict upon users in the earlier stages. On the other hand, as shown by our experiments, a significant loss is incurred if an offline policy is used without any exploration.

In this paper, we propose a remedy for this cold-start problem in tabular OLTR, also called *stochastic ranking bandits* (Zoghi et al., 2017; Lattimore et al., 2018). We show that prior knowledge on the quality of ranked items can greatly improve the efficacy of ranking bandit algorithms, both in theory and practice. There are at least two methods for acquiring this prior: unbiased estimation and offline models. An extensive body of work exists on unbiased estimation of the quality of ranked items, either using click models trained on logged data (Chuklin et al., 2015), counterfactual evaluation (Li et al., 2015; Zoghi et al., 2016; Li et al., 2018), or a combination of these (Dudik et al., 2014). On the other hand, there have been great advances in the quality of offline LTR algorithms (Qin et al., 2021; Burges, 2010), raising the question of whether their predictions could be used as priors for OLTR.

This work makes algorithmic, theoretical, and empirical contributions. First, we adapt two algorithms to OLTR, where the prior distributions on the utilities of items are inputs: BayesUCB (Kaufmann et al., 2012) and *Thompson sampling* (TS) (Thompson, 1933; Chapelle and Li, 2012; Russo et al., 2018). Our work is the first to apply BayesUCB to OLTR. While TS has been applied to OLTR before (Zong et al., 2016; Cheung et al., 2019), we are the first to apply it to the *dependent click model (DCM)* (Guo et al., 2009b), a variant of the *cascade model (CM)* (Richardson et al., 2007; Craswell et al., 2008) that accounts for both the item and position bias.

Second, we derive novel prior-dependent Bayes regret bounds for both BayesUCB and TS in three click models (Chuklin et al., 2015). The main challenges in our derivations are an exponentially large arm space and non-linear objectives (Theorem 2 and Corollary 3).

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

*The work started while being at Google Research.

Our bounds are the first in OLTR that capture how informative the prior is, which can be also observed empirically (Section 7). The novelty of our bounds is discussed more after Theorem 1 and in Section 8.

Finally, we evaluate our algorithms empirically with the following observations: BayesUCB and TS outperform state-of-the-art baselines; our prior-dependent regret bounds are practical and reflect the decrease in regret due to informative priors; and BayesUCB and TS can be integrated with a state-of-the-art LTR system (Qin et al., 2021). This combined approach is evaluated on thousands of queries with more than a hundred items per query.

2 SETTING

We start with introducing some notation. We define $[K] = \{1, \dots, K\}$. For any vectors $u \in \mathbb{R}^d$ and $v \in [d]^n$, we let $u(v) \in \mathbb{R}^n$ be a vector whose i-th entry is the v_i-th entry of u, for any $i \in [n]$. We treat vectors as sets when needed. We use $\tilde{O}$ for the big O notation up to logarithmic factors.

We have a ranking problem, with L ground items and display a ranked list of K items. We denote the set of all tuples of K distinct items out of L by $\Pi_K(L)$ and call $I \in \Pi_K(L)$ a *list* of items. Each item has some utility, which we call an attraction probability. We denote the *attraction probability* of item i by $\theta_i \in [0, 1]$ and let $\theta = (\theta_i)_{i=1}^L$.

We formalize our OLTR problem as follows. The agent interacts with an environment, such as users in a recommender system, over n rounds. In round $t \in [n]$, the *attraction* of each item i is drawn i.i.d. as $Y_{i,t} \sim \mathrm{Ber}(\theta_i)$, across both the items and rounds. The attractions can be viewed as *unknown realized preferences* of the user in round t. The agent recommends a list of items $I_t \in \Pi_K(L)$ and the user *examines* it sequentially from the first item. When the item is examined and attractive, the user *clicks* on it. Whether the user is satisfied with the item, or proceeds to the next item, depends on the model of user interaction. Such models are known as *click models* (Chuklin et al., 2015) and we examine several of them below.

Let $Y_{I,t} = (Y_{I(k),t})_{k=1}^K$ be a vector of realized attractions of all items in list I in round t. We define the satisfaction with list I in round t, or *reward*, as $f(Y_{I,t})$, where $f : [0, 1]^K \to \mathbb{R}_{\geq 0}$ is a *reward function* to be defined precisely later and $\mathbb{R}_{\geq 0}$ is the set of non-negative real numbers. The *mean reward of list* I is

$$\mu_I = \mathbb{E}\left[f(Y_{I,t})\right] = \mathbb{E}_{Y' \sim \prod_{i \in I} \mathrm{Ber}(\cdot; \theta_i)}\left[f(Y')\right].$$

While this quantity depends on θ and list I, it is independent of the round t because the attractions of items are drawn i.i.d. across the rounds. The *optimal list* is $I_* = \arg\max_{I \in \Pi_K(L)} \mu_I$ and depends on θ.

Our objective is to maximize the satisfaction of users with recommended lists over n rounds. We formalize this problem as regret minimization. Let the expected regret in round t be $R_t = \mu_{I_*} - \mu_{I_t}$. Our goal is to minimize its sum over n rounds, $R(n) = \mathbb{E}\left[\sum_{t=1}^n R_t\right]$, where the expectation is taken over the randomness in the algorithm, attractions $Y_{i,t}$, and attraction probabilities $\theta \sim P_0$, where P_0 is the *prior distribution* over attraction probabilities known by the learning agent. The quantity $R(n)$ is called the *Bayes regret* (Russo and Van Roy, 2014). We use it as a metric because it naturally captures the notion of side information P_0, and can be utilized in both the algorithm design and analysis (Russo and Van Roy, 2014). This would not be possible with the so-called frequentist regret, where θ is unknown but fixed.

The prior is factored as $P_0(\theta) = \prod_{i=1}^L \mathrm{Beta}(\theta_i; \alpha_i, \beta_i)$, where α_i and β_i are the parameters of the beta prior for θ_i. We use the beta prior because it is a conjugate prior to clicks, which are our observations. The beta prior is also universal. Roughly speaking, $\mathrm{Beta}(\alpha, \beta)$ can be viewed as a Gaussian with mean $\alpha/(\alpha + \beta)$ and variance $\alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$ on $[0, 1]$, and has a comparable representational power.

3 ALGORITHMS

We present two Bayesian algorithms for OLTR, which are motivated by BayesUCB (Kaufmann et al., 2012) and Thompson sampling (Thompson, 1933; Chapelle and Li, 2012; Russo et al., 2018). While Thompson sampling has been applied to OLTR before (Zong et al., 2016; Cheung et al., 2019), our work is the first application of BayesUCB to OLTR. We call both proposed algorithms by the name of the base algorithm, BayesUCB and TS, since it is clear from context that they are applied to OLTR in this work.

The key idea in our algorithms is to maintain a posterior distribution over the attraction probabilities of all items and then act optimistically with respect to it. By the Bayes rule, the posterior in round t is

$$P_t(\theta) = \prod_{i=1}^L \mathrm{Beta}(\theta_i; \alpha_{i,t}, \beta_{i,t}), \tag{1}$$

$$\alpha_{i,t} = \alpha_i + \sum_{\ell=1}^{t-1} O_{i,\ell} Y_{i,\ell},$$

$$\beta_{i,t} = \beta_i + \sum_{\ell=1}^{t-1} O_{i,\ell}(1 - Y_{i,\ell}),$$

where $O_{i,\ell} \in \{0, 1\}$ is an indicator that the attraction

Algorithm 1 BayesUCB for click models.

```
$P_1 \leftarrow P_0$
for $t = 1, \dots, n$ do
    Compute UCBs $U_t = (U_{i,t})_{i=1}^L$
    $I_t \leftarrow$ Top $K$ items with respect to $U_t$
    Recommend $I_t$ and get click feedback
    Update posterior $P_{t+1}$
```

Algorithm 2 TS for click models.

```
$P_1 \leftarrow P_0$
for $t = 1, \dots, n$ do
    Sample $\theta_t = (\theta_{i,t})_{i=1}^L \sim P_t$
    $I_t \leftarrow$ Top $K$ items with respect to $\theta_t$
    Recommend $I_t$ and get click feedback
    Update posterior $P_{t+1}$
```

of item i is observed in round ℓ. Roughly speaking, $\alpha_{i,t}$ is the number of clicks on item i up to round t increased by prior pseudo-counts α_i, and analogously $\beta_{i,t}$ is the number of observed "no clicks". Estimating if the item is observed, $O_{i,\ell} = 1$, is a hard problem. However, in the click models studied in this work, it can be done for a subset of recommended items, which is sufficient to design efficient learning algorithms. We precisely define $O_{i,\ell}$ when we study these models.

Our algorithms are presented in Algorithms 1 and 2, and work as follows. In round t, they recommend K most attractive items I_t, in descending order of per-item statistics derived from the posterior P_t. Greedily choosing most attractive items is a popular design in OLTR (Kveton et al., 2015a; Katariya et al., 2016), and can be extended to both context (Zong et al., 2016; Li et al., 2016) and diversity (Hiranandani et al., 2019). The per-item statistics in BayesUCB are upper confidence bounds. The *upper confidence bound (UCB)* of item i in round t is

$$U_{i,t} = \min \left\{ c \in [0,1] : \int_{y=c}^{1} \mathrm{Beta}(y; \alpha_{i,t}, \beta_{i,t}) \, \mathrm{d}y \leq \delta \right\} .$$

This is the lowest value c such that the probability of event $\theta_i \geq c$ is at most δ, where $\delta > 0$ is a tunable parameter. Note that $U_{i,t}$ is in $[0,1]$. The per-item statistics in TS are *posterior-sampled attraction probabilities* $\theta_t \sim P_t$. After the list I_t is recommended, both algorithms observe clicks and update their posteriors P_{t+1}, as defined in (1). How the clicks relate to the attractions of items depends on a given click model, which we discuss next.

4 DOCUMENT-BASED CLICK MODEL

In this section, we derive regret bounds for BayesUCB and TS in a simple click model, which sets stage for more practical models in the following sections.

4.1 Click Model

The *document-based click model (DCTR)* (Craswell et al., 2008; Chuklin et al., 2015) is one of the simplest models of how a user interacts with a ranked list of items I_t. The model is defined as follows. The user examines all positions in I_t. If the item at position k is attractive, $Y_{I_t(k),t} = 1$, the user clicks on it.

A natural notion of reward in this model is the number of clicks in a list. For any list I, this reward and its expectation can be written as

$$f(Y_{I,t}) = \sum_{k=1}^{K} Y_{I(k),t} \,, \quad \mu_I = \mathbb{E}\left[f(Y_{I,t}) \,|\, \theta \right] = \sum_{k=1}^{K} \theta_{I(k)} \,.$$

This objective is maximized by choosing the K most attractive items. This model of reward and feedback is known in online learning as *semi-bandits* (Gai et al., 2012; Chen et al., 2013; Kveton et al., 2014, 2015b).

4.2 Regret Bound

We derive a regret bound for BayesUCB and TS for any failure probability $\delta > 0$. Proofs of all technical lemmas are deferred to Appendix A.

Theorem 1. *The regret of* BayesUCB *and* TS *in DCTR is bounded with probability* $1 - \delta$ *as* $R(n) \leq$

$$\sqrt{2KLn \log(1/\delta)} \sqrt{\log \left(1 + \frac{n}{L} \sum_{i=1}^{L} \frac{1}{\alpha_i + \beta_i} \right)} + 2L\delta n \,.$$

For $\delta = 1/n$, the bound is $O(\sqrt{KLn} \log n)$ and sublinear in the horizon n. The novelty of this result is in its dependence on prior, in the second logarithmic term above. Since $\sqrt{\log x}$ is an increasing function of x, the regret decreases when $1/(\alpha_i + \beta_i)$ does for any i. By Lemma 4 (Appendix A), this quantity is a scaled sub-Gaussianity parameter of the prior distribution of item i, which we call a *prior width*. Therefore, the regret decreases whenever the prior width of any item decreases, and we become more certain about its attraction probability. Since $\log(1 + x) \to 0$ as $x \to 0$, the regret approaches zero when $\alpha_i + \beta_i \to \infty$ for all $i \in [L]$. In this case, the problem instance becomes increasingly more certain, and no exploration is needed to find the optimal list.

Our bound approaches infinity when $\alpha_i, \beta_i \to 0$. This can be fixed by recommending all items once initially. The extra regret for this initialization is $O(L)$ and constant in n. After the initialization, we get priors that satisfy $\alpha_i + \beta_i \geq 1$.

A Bayes regret lower bound exists for a K-armed bandit (Lai, 1987). However, it is unclear how to generalize it to structured problems, and even seminal works on Bayes regret minimization do not match it (Russo and Van Roy, 2014, 2016). Therefore, we validate the tightness of our bounds empirically (Section 7). Our experiments show that the bounds are just an order of magnitude off for a wide range of prior widths, which is tight relative to other regret bounds derived in the literature.

Our analysis can be extended to misspecified priors. Roughly speaking, when α_i and β_i are ε-close to the true values, we would get an $O(\varepsilon n^2)$ extra regret. This claim can be proved by extending Lemma 5 in Kveton et al. (2021), which shows in a Gaussian bandit that the extra regret is $O(\varepsilon n^2)$ when the prior mean of all arms is misspecified by $O(\varepsilon)$. The extension is possible because the beta distribution is sub-Gaussian with variance proxy $0.25/(\alpha + \beta + 1)$.

4.3 Proof of Theorem 1

Before starting, we define some notation. The posterior attraction probability and the corresponding confidence interval width for any item i in round t are

$$\hat{\theta}_{i,t} = \frac{\alpha_{i,t}}{\alpha_{i,t} + \beta_{i,t}}\,, \quad C_{i,t} = U_{i,t} - \hat{\theta}_{i,t}\,. \tag{2}$$

For any list I in round t, $U_{I,t} = \sum_{i \in I} U_{i,t}$ is the sum of its UCBs, $C_{I,t} = \sum_{i \in I} C_{i,t}$ is the sum of its posterior widths, and $\hat{\mu}_{I,t} = \sum_{i \in I} \hat{\theta}_{i,t}$ is the sum of its posterior means.

`BayesUCB` **analysis:** We start with decomposing the regret by conditioning on history H_t, of all actions and observations that define the posterior in (1). To simplify notation, we introduce $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid H_t]$ and decompose the regret as $R(n) = \sum_{t=1}^n \mathbb{E}[\mathbb{E}[R_t \mid H_t]]$. In round t,

$$\begin{aligned} \mathbb{E}_t[R_t] &= \mathbb{E}_t[\mu_{I_*} - \mu_{I_t}] \\ &\leq \mathbb{E}_t[\mu_{I_*} - U_{I_*,t} + U_{I_t,t} - \mu_{I_t}] \\ &= \mathbb{E}_t[\mu_{I_*} - U_{I_*,t}] + \mathbb{E}_t[U_{I_t,t} - \mu_{I_t}]\,. \end{aligned} \tag{3}$$

The inequality holds because `BayesUCB` is optimistic, meaning that $U_{I_t,t} \geq U_{I_*,t}$ conditioned on history H_t. Now we bound both terms above. Let

$$E_t = \left\{\forall i \in [L] : |\theta_i - \hat{\theta}_{i,t}| \leq C_{i,t}\right\} \tag{4}$$

be the event that all confidence intervals in round t hold. For the first term in (3), we have

$$\begin{aligned} \mathbb{E}_t[\mu_{I_*} - U_{I_*,t}] &= \mathbb{E}_t[\mu_{I_*} - \hat{\mu}_{I_*,t}] - \mathbb{E}_t[C_{I_*,t}] \\ &\leq \mathbb{E}_t\left[(\mu_{I_*} - \hat{\mu}_{I_*,t})\mathbb{1}\{\bar{E}_t\}\right]\,. \end{aligned} \tag{5}$$

The inequality follows from $\mathbb{E}_t[(\mu_{I_*} - \hat{\mu}_{I_*,t})\mathbb{1}\{E_t\}] \leq \mathbb{E}_t[C_{I_*,t}]$. From the union bound over all L items and that $C_{i,t}$ is a $(1-\delta)$-probability confidence interval,

$$\mathbb{E}_t\left[(\mu_{I_*} - \hat{\mu}_{I_*,t})\mathbb{1}\{\bar{E}_t\}\right] \leq L\delta\,. \tag{6}$$

An application of the above inequality to (5) leads to $\mathbb{E}_t[\mu_{I_*} - U_{I_*,t}] \leq L\delta$. An analogous line of reasoning leads to an upper bound on the second term in (3),

$$\begin{aligned} \mathbb{E}_t[U_{I_t,t} - \mu_{I_t}] &\leq 2\mathbb{E}_t[C_{I_t,t}] + \mathbb{E}_t\left[(\hat{\mu}_{I_t,t} - \mu_{I_t})\mathbb{1}\{\bar{E}_t\}\right] \\ &\leq 2\mathbb{E}_t[C_{I_t,t}] + L\delta\,. \end{aligned}$$

Next we apply both derived upper bounds to (3) and get $R(n) \leq 2\mathbb{E}\left[\sum_{t=1}^n C_{I_t,t}\right] + 2L\delta n$.

The last step is an upper bound on the sum of confidence interval widths $C_{I_t,t}$. First, note that $C_{I_t,t}$ is a sum of K individual confidence interval widths of items $i \in I_t$. Then, by the Cauchy-Schwarz inequality, we have

$$\sum_{t=1}^n C_{I_t,t} \leq \sqrt{Kn}\sqrt{\sum_{i=1}^L \sum_{t=1}^n \mathbb{1}\{i \in I_t\}\, C_{i,t}^2}\,. \tag{7}$$

By Lemma 4 (Appendix A), we get that $\theta_i - \hat{\theta}_{i,t} \mid H_t$ is σ^2-sub-Gaussian for $\sigma^2 = 0.25/(\alpha_i + \beta_i + T_{i,t} + 1)$, where $T_{i,t} = \sum_{\ell=1}^{t-1} O_{i,\ell}$ denotes the number of times that item i is observed up to round t. As $C_{i,t}$ is the exact confidence interval that fails with probability δ, any other confidence interval must be at least as wide for any fixed δ. This is also true for the corresponding sub-Gaussian interval, and thus

$$C_{i,t} \leq \sqrt{\frac{\log(1/\delta)}{2(\alpha_{i,t} + \beta_{i,t} + 1)}} = \sqrt{\frac{\log(1/\delta)}{2(\alpha_i + \beta_i + T_{i,t} + 1)}}\,.$$

In addition, since no item is displayed more than once per round, we have for any item i that

$$\begin{aligned} \sum_{t=1}^n \mathbb{1}\{i \in I_t\}\, C_{i,t}^2 &\leq \frac{\log(1/\delta)}{2} \sum_{s=1}^n \frac{1}{\alpha_i + \beta_i + s} \\ &\leq \frac{\log(1/\delta)}{2} \log\left(1 + \frac{n}{\alpha_i + \beta_i}\right)\,, \end{aligned}$$

where the last step is by Lemma 5 (Appendix A). Now we use the concavity of the logarithm and get

$$\sum_{i=1}^L \log\left(1 + \frac{n}{\alpha_i + \beta_i}\right) \leq L \log\left(1 + \frac{n}{L} \sum_{i=1}^L \frac{1}{\alpha_i + \beta_i}\right)\,.$$

Finally, we chain all inequalities and get the claim.

Thompson sampling analysis: The only difference is that the inequality in (3) becomes an equality. The reason is that the upper confidence bound is a deterministic function of H_t, and that I_* and I_t are i.i.d. conditioned on H_t. The latter is by the design of `TS`.

5 CASCADE MODEL

The *cascade model (CM)* (Richardson et al., 2007; Craswell et al., 2008) is a popular click model that captures item bias, in that lower ranked items are less likely to be clicked due to higher ranked items. The model is defined as follows.

5.1 Click Model

The user examines the first position in I_t with probability 1. If position k is examined and the item at that position is attractive, $Y_{I_t(k),t} = 1$, the user clicks on it and does not examine any other position. If the user does not click, they examine the next position $k + 1$.

A natural notion of reward in this model is a click on the list (Kveton et al., 2015a; Zong et al., 2016; Li et al., 2016), an indicator that at least one item in the list is attractive. For any list I, this reward and its expectation can be written as

$$f(Y_{I,t}) = 1 - \prod_{k=1}^{K} (1 - Y_{I(k),t})\,,$$

$$\mu_I = \mathbb{E}\left[f(Y_{I,t}) \,|\, \theta\right] = 1 - \prod_{k=1}^{K} (1 - \theta_{I(k)})\,,$$

where the last equality follows from the independence of attractions (Section 2). As in Section 4, μ_I is maximized by the K most attractive items.

The main difference from Section 4 is that the user exits upon a click. Therefore, the item attractions are only observed up to that click. We formally define the clicked position as $S_t = \min\left\{k \in [K] : Y_{I_t(k),t} = 1\right\}$, where $\min \emptyset = K$. Thus $S_t = K$ when the user does not click on any item. Using this definition, we have that $O_{i,t} = \sum_{k=1}^{S_t} \mathbb{1}\{i = I_t(k)\}$ is the indicator of observing item i in round t.

5.2 Regret Bound

We derive a regret bound for `BayesUCB` and `TS` for any failure probability δ. It is sublinear in n for $\delta = 1/n$.

Theorem 2. *The regret of* `BayesUCB` *and* `TS` *in CM is bounded with probability* $1 - \delta$ *as* $R(n) \leq$

$$\sqrt{2KLn\log(1/\delta)}\sqrt{\log\left(1 + \frac{n}{L}\sum_{i=1}^{L}\frac{1}{\alpha_i + \beta_i}\right)} + 2L\delta n\,.$$

The above bound is the same as Theorem 1. This is surprising, as the learning agent gets less feedback in the CM than in the DCTR; and so we would expect a higher regret. This does not happen since the range of rewards also changed, from $[0, K]$ to $[0, 1]$. Theorem 2 also cannot significantly improve upon Theorem 1 because the CM behaves similarly to the DCTR when the item attractions are low. In this case, for any list I, $1 - \prod_{k=1}^{K}(1 - \theta_{I(k)}) \approx \sum_{k=1}^{K}\theta_{I(k)}$ and all items in I are likely to be examined. We compare this result to prior works in Section 8.

Theorem 2 is proved using a novel regret decomposition, which differs from prior works. In Kveton et al. (2015a), the optimal and recommended lists in round t are fixed given the history. Under this assumption, there exists a deterministic regret decomposition of a recommended list into the regret of individual items in it (Lemma 1 therein). This cannot be done here. In `BayesUCB`, the optimal list is random given history. In `TS`, both the optimal and recommended lists are random given history. This means that the regret decomposition is random and a different approach is needed. To address this issue, we carefully introduce an upper bound on the probability of clicking on a list, which is a deterministic function of the history, and propagate it through the regret decomposition. This requires a two-sided version of Lemma 1 in Kveton et al. (2015a), which is Lemma 7 in Appendix A.

5.3 Proof of Theorem 2

`BayesUCB` analysis: To simplify notation, we define $U_{I,t} = 1 - \prod_{i \in I}(1 - U_{i,t})$. Our first step is the regret decomposition in (3). Then we bound $\mathbb{E}_t\left[\mu_{I_*} - U_{I_*,t}\right]$ and $\mathbb{E}_t\left[U_{I_t,t} - \mu_{I_t}\right]$ as follows. We start with an upper bound on

$$\begin{aligned}\mu_{I_*} - U_{I_*,t} &= \prod_{i \in I_*}(1 - U_{i,t}) - \prod_{i \in I_*}(1 - \theta_i) \\ &= \sum_{k=1}^{K}\left(\prod_{j=1}^{k-1}(1 - \theta_{I_*(j)})\right)(\theta_{I_*(k)} - U_{I_*(k),t}) \\ &\quad \times \left(\prod_{j=k+1}^{K}(1 - U_{I_*(j),t})\right),\end{aligned}$$

where the last equality is from the second claim in Lemma 7 (Appendix A). Note that $1 - \theta_{I_*(j)}$ and $1 - U_{I_*(j),t}$ are in $[0, 1]$. Moreover, when event E_t in (4) occurs, $\theta_{I_*(k)} - U_{I_*(k),t} \leq 0$ for all $k \in [K]$. Hence we get $\mu_{I_*} - U_{I_*,t} \leq 0$ on event E_t. On the other hand, as in (6), event $\bar{E}_t$ occurs with probability at most $L\delta$. It follows that

$$\begin{aligned}\mathbb{E}_t\left[\mu_{I_*} - U_{I_*,t}\right] = \mathbb{E}_t\left[(\mu_{I_*} - U_{I_*,t})\mathbb{1}\{E_t\}\right] + \\ \mathbb{E}_t\left[(\mu_{I_*} - U_{I_*,t})\mathbb{1}\left\{\bar{E}_t\right\}\right] \leq L\delta\,.\end{aligned}$$

We continue with an upper bound on

$$\begin{aligned}U_{I_t,t} - \mu_{I_t} &= \prod_{i \in I_t}(1 - \theta_i) - \prod_{i \in I_t}(1 - U_{i,t}) \\ &= \sum_{k=1}^{K}\left(\prod_{j=1}^{k-1}(1 - \theta_{I_t(j)})\right)(U_{I_t(k),t} - \theta_{I_t(k)}) \\ &\quad \times \left(\prod_{j=k+1}^{K}(1 - U_{I_t(j),t})\right),\end{aligned}$$

where the last equality follows from the first claim in Lemma 7 (Appendix A). We note again that $1 - \theta_{I_t(j)}$ and $1 - U_{I_t(j),t}$ are in $[0,1]$. Moreover, when event E_t occurs, we have that $U_{I_t(k),t} - \theta_{I_t(k)} \leq 2C_{I_t(k),t}$ holds for all $k \in [K]$. Finally, note that $\prod_{j=1}^{k-1}(1 - \theta_{I_t(j)})$ is the probability that item $I_t(k)$, at position k in I_t, is examined. Let $O_{I_t(k),t}$ be an indicator of this event, as defined in Section 5. Then, based on these facts, we have on event E_t that

$$\mathbb{E}_t\left[(U_{I_t,t} - \mu_{I_t})\mathbb{1}\{E_t\}\right] \leq 2\mathbb{E}_t\left[\sum_{k=1}^{K} O_{I_t(k),t}C_{I_t(k),t}\right].$$

Since event $\bar{E}_t$ occurs with probability at most $L\delta$, it follows that

$$\begin{aligned}\mathbb{E}_t\left[U_{I_t,t} - \mu_{I_t}\right] = \mathbb{E}_t\left[(U_{I_t,t} - \mu_{I_t})\mathbb{1}\{E_t\}\right] + \\ \mathbb{E}_t\left[(U_{I_t,t} - \mu_{I_t})\mathbb{1}\left\{\bar{E}_t\right\}\right] \\ \leq 2\mathbb{E}_t\left[\sum_{k=1}^{K} O_{I_t(k),t}C_{I_t(k),t}\right] + L\delta\,.\end{aligned}$$

Now we apply the above bounds to (3) and get $R(n) \leq 2\mathbb{E}\left[\sum_{t=1}^{n}\sum_{k=1}^{K} O_{I_t(k),t}C_{I_t(k),t}\right] + 2L\delta n$. The last step is an upper bound on the sum of confidence interval widths $C_{I_t(k),t}$. This part of the proof is identical to Theorem 1. This is because all items could have been examined and contribute to the regret, proportionally to their confidence interval widths.

Thompson sampling analysis: This is an adaptation of the UCB bound, where one inequality is replaced with an equality, as in Section 4.3.

6 DEPENDENT CLICK MODEL

The *dependent click model (DCM)* (Guo et al., 2009b) is a cascade-like model that also captures position bias, in that lower ranked items are less likely to be satisfactory. The model has K additional parameters, one *satisfaction probability* v_k for each position k. We let $v = (v_k)_{k=1}^{K}$. In round t, the *satisfaction* of each position k is drawn i.i.d. as $V_{k,t} \sim \text{Ber}(v_k)$, independently of all other attractions and satisfactions in any round. The model is defined as follows.

6.1 Click Model

The user examines the first position in I_t with probability 1. If position k is examined and the item at that position is attractive, $Y_{I_t(k),t} = 1$, the user clicks on it. Upon the click, the user is *satisfied* when $V_{k,t} = 1$ and stops examining the remaining items. If the user does not click or is not satisfied, they examine the next position $k+1$. Because of this, the DCM permits multiple clicks.

A natural reward in this model is satisfaction with the list (Katariya et al., 2016), that the user leaves satisfied upon a click. Note that this is unobserved, as we do not know whether the last click is satisfactory. Nonetheless, for any list I, this reward and its expectation can be written as

$$f(Y_{I,t}) = 1 - \prod_{k=1}^{K}(1 - V_{k,t}Y_{I(k),t})\,,$$

$$\mu_I = \mathbb{E}\left[f(Y_{I,t}) \,|\, \theta, v\right] = 1 - \prod_{k=1}^{K}(1 - v_k\theta_{I(k)})\,,$$

where the last equality follows from the independence of $Y_{i,t}$ and $V_{k,t}$. As in Section 5, μ_I is maximized by the K most attractive items (Katariya et al., 2016), when $v_1 \geq \cdots \geq v_K$. Thus the satisfaction probabilities do not have to be known to identify the optimal list I_*.

The main difference from Section 5 is that the user exits upon a satisfactory click. Although we do not know whether a click is satisfactory, we know that the user does not click after leaving satisfied. Therefore, the item attractions are guaranteed to be observed up to the last clicked position, which we denote by S_t. As in Section 5, we set $S_t = K$ when the user does not click on any item. Using this definition, we have that $O_{i,t} = \sum_{k=1}^{S_t}\mathbb{1}\{i = I_t(k)\}$ is the indicator of observing item i in round t.

6.2 Regret Bound

Our regret bound for `BayesUCB` and `TS`, which is sublinear in n for $\delta = 1/n$, is presented below.

Corollary 3. *Let*

$$\theta_{\text{UCB}} = \max\left\{\tfrac{\alpha_i}{\alpha_i+\beta_i} + \sqrt{\tfrac{\log n}{2(\alpha_i+\beta_i+1)}}\right\}_{i \in [L]} < 1$$

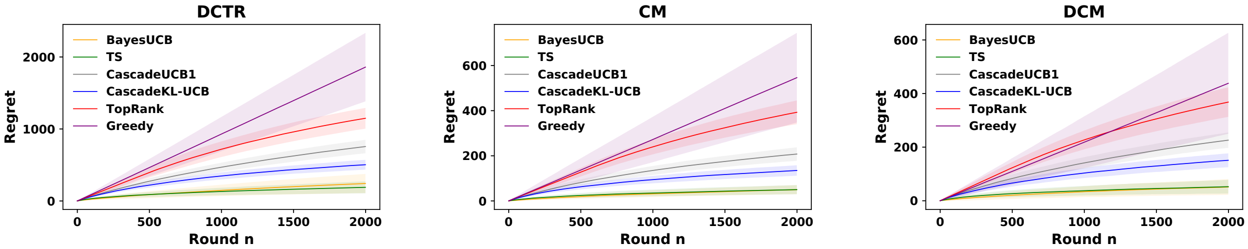

Figure 1: Comparison of BayesUCB and TS to four baselines in three click models.

and $\eta_{\text{UCB}} = (1 - \theta_{\text{UCB}})^{K-1}$. *Then the regret of* BayesUCB *and* TS *in DCM is bounded with probability* $1 - \delta$ *as*

$$R(n) \leq \eta_{\text{UCB}}^{-1} \left[\sum_{k=1}^{K} (v_k - v_{k+1}) \sqrt{2kLn \log(1/\delta)} \times \sqrt{\log\left(1 + \frac{n}{L} \sum_{i=1}^{L} \frac{1}{\alpha_i + \beta_i}\right)} + 2L\delta n \right] + L\,.$$

Proof. The bound is based on the reduction in Theorem 2 of Katariya et al. (2016), which says the following. Let $1 \geq v_1 \geq \cdots \geq v_{K+1} = 0$ and $R_k(n)$ be the regret of any bandit algorithm in the CM on a list of length k. Then $R(n) \leq \eta^{-1} \sum_{k=1}^{K} (v_k - v_{k+1}) R_k(n)$, where $\eta = (1 - \theta_{\max})^{K-1}$ and $\theta_{\max} = \max_{i \in [L]} \theta_i$. In our case, $\theta_{\max}$ is random because θ is. Therefore, we replace it with a high-probability upper bound, which holds with probability L/n. The bound follows from the sub-Gaussianity of prior P_0, which is a product of betas (Lemma 4 in Appendix A). □

Compared to Theorem 2, the bound depends on differences of the consecutive satisfaction probabilities. The maximum is attained at $v_1 = \cdots = v_K = 1$ and $v_{K+1} = 0$. In this case, the bound reduces to that in Theorem 2, up to the factor η_{UCB}^{-1}. In all other cases, the bound is lower and indicates that these cases are easier. The extra factor of η_{UCB}^{-1} is not expected to be large. The reason is that θ_{UCB} tends to be small, since click probabilities tend to be low.

7 EXPERIMENTS

We conduct both synthetic and more realistic experiments, where the priors are generated using offline-trained models and might not exactly match the click distribution. Our code is included in the supplementary material, and information about computational resources is provided in Appendix C.

7.1 Synthetic Experiments

In the synthetic experiments, we generate problem instances by sampling them from a prior. The attraction probability of item i is sampled as $\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$ and its realization in round t is $Y_{i,t} \sim \text{Ber}(\theta_i)$. The clicks are simulated based on $Y_{i,t}$ and click models in Sections 4 to 6. The algorithms know the prior but not θ_i. Such knowledge gives any algorithm that uses it an advantage. This raises a natural question of prior misspecification, which we investigate as well.

We compare BayesUCB and TS (Algorithms 1 and 2) to two types of baselines. The first type are state-of-the-art adaptive algorithms that do not use the prior: TopRank (Lattimore et al., 2018); and CascadeUCB1 and CascadeKL-UCB (Kveton et al., 2015a). The second type is a non-adaptive baseline where the items are ranked according to their maximum-a-posteriori attraction probabilities estimated from the same prior as in BayesUCB and TS. We call it Greedy.

The number of items is $L = 30$ and $K = 3$. We set $\beta_i = 10$ and sample α_i from $[10]$ uniformly at random, for all items i. This generates mean attraction probabilities between 0.09 and 0.5. So no item is overly attractive. We sample $(\alpha_i)_{i=1}^{K}$ for 20 times, and for each we sample the attraction probability of items as $\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$ for 20 times. This yields 400 bandit instances. Each click model and algorithm are simulated on these instances for $n = 2000$ rounds. In the DCM, we set $v_1 = \cdots = v_K = 0.5$. Figure 1 shows the n-round regret for all algorithms and click models. We observe that BayesUCB and TS perform similarly, and significantly outperform the baselines.

7.2 Effect of the Prior

To study how the prior affects regret, we vary it from less informative to more informative. Specifically, we vary $\gamma \in \{10, 20, 50, 100, 200, 500, 1000\}$, and then set $\alpha_i = \gamma$ and $\beta_i = 10\gamma$ for all items i. Small values ($\gamma = 10$) correspond to a wide and uninformative prior, while large values ($\gamma = 1000$) correspond to a narrow and informative prior. We sample $\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$

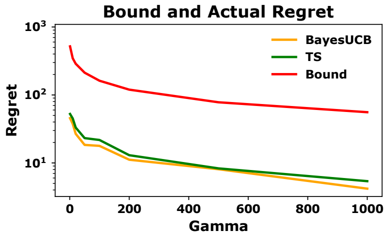

Figure 2: Actual regret and its upper bound (Theorem 2) for varying prior width in the cascade model.

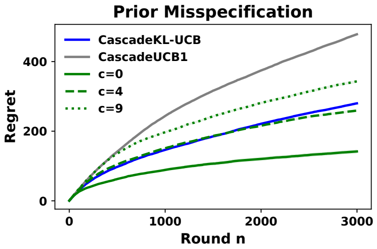

Figure 3: Regret of TS with various prior misspecifications c.

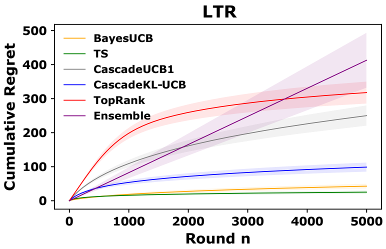

Figure 4: Comparison of BayesUCB and TS to four baselines in the real-world experiment.

for 100 times, and run BayesUCB and TS for $n = 500$, $L = 30$, and $K = 3$ in the cascade model. Figure 2 shows the n-round regret at $n = 500$ as a function of γ. We also report the regret bound from Theorem 2. We observe that the regret decreases as the prior becomes more informative. Our bound is fairly tight, within a factor of 10, and nicely matches the decrease in the regret. Additional results with varying attraction probabilities are presented in Appendix B.

7.3 Prior Misspecification

Now we study prior misspecification. The true prior is $\mathrm{Beta}(1, 10)$ but the prior in TS is $\mathrm{Beta}(1 + c, 10 - c)$, with $c \in [0, 9]$. Therefore, when $c > 0$, the prior in TS is misspecified. We use the same experimental setting as in Section 7.2, except that $n = 3000$. In Figure 3, we show the n-round regret of TS for various levels of misspecification $c \in \{0, 4, 9\}$. Although the regret increases with c, it is well behaved and flattens even for large c. For reference, we also show CascadeUCB1 and CascadeKL-UCB. We observe that CascadeKL-UCB outperforms TS only when the prior is significantly misspecified ($c = 9$), while CascadeUCB1 is not competitive. Results for BayesUCB are similar and we do not report them.

7.4 Real-World Experiment

We experiment with the Microsoft Learning to Rank Web30K dataset[1] (Qin and Liu, 2013), which contains 18,919 training and 6,306 test queries, with an average of $L = 120$ documents per query. To generate our priors, we train $m = 10$ LTR models, which output m scores for each query-document pair. Each model is trained on 90% of randomly sampled training data without replacement. We use a state-of-the-art LTR model (Qin et al., 2021) based on neural networks, which outperforms gradient-boosted decision trees on large-scale LTR benchmarks. To guarantee that the scores are in $[0, 1]$ both during training and inference, we employ the sigmoid cross-entropy loss (Pasumarthi et al., 2019). For training, the ground-truth relevance scores $[0, 1, 2, 3, 4]$ are mapped to $[0, 0.25, 0.5, 0.75, 1]$. For each query-document pair in the test set, we generate m scores $s_1, \dots, s_m$ and compute the beta prior as $\mathrm{Beta}\left(\sum_{i=1}^m s_i, \sum_{i=1}^m (1 - s_i)\right)$. This prior uses each score as a soft vote for the attractiveness of an item, and it is used by both BayesUCB and TS.

To simulate clicks in evaluation, we map the ground-truth relevance scores $[0, 1, 2, 3, 4]$ to attraction probabilities $[0, 0.2, 0.4, 0.8, 1]$ according to “perfect mapping” (Table 2 in Hofmann et al. (2013)). This mapping differs from the one used in training, and represents a typical mismatch between offline labels and the ground truth. This is how we introduce prior misspecification in this experiment. During evaluation, for each of 6,306 test queries, we simulate each algorithm for $n = 5000$ rounds with $K = 10$, and repeat this 50 times. This results in more than 300k bandit instances in this experiment. Figure 4 shows the n-round regret of all algorithms. The results are similar to Figure 1, showing that our prior-based algorithms significantly outperform the baselines. We also evaluate algorithm Ensemble, which ranks items according to their predicted scores $s = \sum_{i=1}^m s_i$. This baseline represents non-adaptive algorithms, which rely solely on offline-trained models and are common in practice. As expected, Ensemble incurs linear regret.

8 RELATED WORK

This section reviews related work. To reduce clutter, all related regret bounds are summarized in Table 1.

The document-based click model (Section 4) was proposed by Craswell et al. (2008). An online learning variant of this model is known as a semi-bandit (Gai et al., 2012; Chen et al., 2013; Kveton et al., 2014, 2015b). Kveton et al. (2015b) analyzed a UCB algorithm for semi-bandits. Their gap-free regret bound has a similar form to Theorem 1 but does not depend

[1] https://www.microsoft.com/en-us/research/project/mslr/

Model	Prior work	Gap-dependent bound	Gap-free bound
DCTR	Kveton et al. (2015b)	$O(KL\Delta^{-1}\log n)$	$O(\sqrt{KLn\log n})$
	Russo and Van Roy (2016)		$O(\sqrt{(L/K)\log(L/K)})$
	Wen et al. (2015)		$\tilde{O}(K\sqrt{d\min\{\log L, d\}\, n})$
CM	Kveton et al. (2015a)	$O((L-K)\Delta^{-1}\log n)$	$O(\sqrt{KLn\log n})$
	Cheung et al. (2019)		$O(\sqrt{KLn}\log n + L\log^{5/2} n)$
DCM	Katariya et al. (2016)	$O((L-K)\Delta^{-1}\log n)$	$O(\sqrt{KLn\log n})$

Table 1: Related regret bounds from prior works. The gap-free bounds of Kveton et al. (2015a) and Katariya et al. (2016) are obtained by trivial reductions.

on prior. Russo and Van Roy (2016) proved a general prior-dependent Bayes regret bound, but did not instantiate it. Wen et al. (2015) proved a Bayes regret bound for contextual semi-bandits with d features. In our case, $d = L$. This bound does not show that the prior is beneficial, because it requires the prior width to be as large as reward noise. Zuo et al. (2020) also derived a prior-independent Bayes regret bound. To the best of our knowledge, Theorem 1 in this work is the first prior-dependent Bayes regret bound for semi-bandits where the effect of the prior width is clearly stated.

The cascade model of user behavior (Section 5) was introduced by Richardson et al. (2007) and Craswell et al. (2008). The first work on OLTR in this model was Kveton et al. (2015a). They proposed UCBs algorithm and analyzed them. Their gap-free bound is similar to Theorem 2 but does not depend on prior. Thompson sampling in the cascade model was proposed by Zong et al. (2016) and Cheung et al. (2019) proved a gap-free bound for it. The bound is frequentist and has a huge constant of $4\sqrt{\pi}e^{8064}$ (Lemma 4.3 and the last equation in Section 4 therein). In contrast, our bound is Bayesian, reflects prior, and does not contain any huge constants. We are the first to derive a Bayes regret bound for this problem class. The dependent click model (Section 6) was proposed by Guo et al. (2009b) and Katariya et al. (2016) used it in OLTR. Their gap-free bound is similar to Corollary 3 but does not depend on prior. This work is the first application of Thompson sampling to the DCM, including analyzing it.

Numerous click models exist (Agichtein et al., 2006; Richardson et al., 2007; Craswell et al., 2008; Guo et al., 2009a,b; Chapelle and Zhang, 2009; Chuklin et al., 2015). General OLTR algorithms with guarantees in multiple click models also exist, such as Zoghi et al. (2016) and Lattimore et al. (2018). The state of the art is TopRank (Lattimore et al., 2018) and it has a gap-free bound of $O(\sqrt{K^3 Ln\log n})$. This bound does not depend on prior and is worse than our bounds by a factor of $O(K)$.

9 CONCLUSIONS

While online learning to rank has been studied extensively, most proposed algorithms cannot take prior information on the utility of ranked items into account. Even if they do, such as in Thompson sampling, that information is not reflected in their regret bounds. We make significant progress on this topic. First, we propose a new prior-dependent algorithm for OLTR based on BayesUCB (Kaufmann et al., 2012). Second, we analyze BayesUCB and Thompson sampling in three click models, and derive their prior-dependent Bayes regret bounds. These bounds show that the regret decreases as the prior narrows and we validate them empirically. These are the first such bounds for OLTR. Finally, we show how BayesUCB and Thompson sampling can be used with a state-of-the-art offline ranker, and evaluate them in simulation on thousands of queries.

Many concepts presented here are general and can be extended to context (Zong et al., 2016; Li et al., 2016), diversity (Hiranandani et al., 2019), and other models of user interaction, such as the position-based model (Craswell et al., 2008). We leave this for future work.

References

E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference*, pages 3–10, 2006.

C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2012.

O. Chapelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web*, pages 1–10, 2009.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework, results and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

W. C. Cheung, V. Tan, and Z. Zhong. A Thompson sampling algorithm for cascading bandits. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 438–447, 2019.

A. Chuklin, I. Markov, and M. de Rijke. *Click Models for Web Search.* Morgan & Claypool Publishers, 2015.

N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining*, pages 87–94, 2008.

M. Dudik, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

K. Falk. *Practical Recommender Systems.* Manning Publications, 2019.

Y. Gai, B. Krishnamachari, and R. Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. M. Wang, and C. Faloutsos. Click chain model in web search. In *Proceedings of the 18th International Conference on World Wide Web*, pages 11–20, 2009a.

F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, pages 124–131, 2009b.

G. Hiranandani, H. Singh, P. Gupta, I. A. Burhanuddin, Z. Wen, and B. Kveton. Cascading linear submodular bandits: Accounting for position bias and diversity in online learning to rank. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, 2019.

K. Hofmann, S. Whiteson, and M. de Rijke. Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Transactions on Information Systems*, 31(4):1–43, 2013.

R. Jagerman, H. Oosterhuis, and M. de Rijke. To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions. In *Proceedings of the 42nd ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

S. Katariya, B. Kveton, C. Szepesvari, and Z. Wen. DCM bandits: Learning to rank with multiple clicks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1215–1224, 2016.

E. Kaufmann, O. Cappe, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 592–600, 2012.

B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 420–429, 2014.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015a.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015b.

B. Kveton, M. Konobeev, M. Zaheer, C.-W. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 1987.

T. Lattimore, B. Kveton, S. Li, and C. Szepesvari. TopRank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems 31*, pages 3949–3958, 2018.

C. Li, B. Kveton, T. Lattimore, I. Markov, M. de Rijke, C. Szepesvari, and M. Zoghi. BubbleRank: Safe online learning to re-rank via implicit click feedback. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, 2019.

L. Li, S. Chen, J. Kleban, and A. Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

S. Li, B. Wang, S. Zhang, and W. Chen. Contextual combinatorial cascading bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1245–1253, 2016.

S. Li, Y. Abbasi-Yadkori, B. Kveton, S. Muthukrishnan, V. Vinay, and Z. Wen. Offline evaluation of ranking policies with click models. In *Proceedings of* *the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1685–1694, 2018.

T.-Y. Liu. *Learning to Rank for Information Retrieval*. Springer Science & Business Media, 2011.

O. Marchal and J. Arbel. On the sub-Gaussianity of the beta and Dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017.

R. K. Pasumarthi, S. Bruch, X. Wang, C. Li, M. Bendersky, M. Najork, J. Pfeifer, N. Golbandi, R. Anil, and S. Wolf. TF-Ranking: Scalable TensorFlow library for learning-to-rank. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2970–2978, 2019.

T. Qin and T.-Y. Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013. URL https://arxiv.org/abs/1306.2597.

Z. Qin, L. Yan, H. Zhuang, Y. Tay, R. K. Pasumarthi, X. Wang, M. Bendersky, and M. Najork. Are neural rankers still outperformed by gradient boosted decision trees? In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pages 784–791, 2008.

M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web*, pages 521–530, 2007.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Y. Tagami, S. Ono, K. Yamamoto, K. Tsukamoto, and A. Tajima. CTR prediction for contextual advertising: Learning-to-rank approach. In *Proceedings of the 7th International Workshop on Data Mining for Online Advertising*, 2013.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

H. Wang, S. Kim, E. McCord-Snook, Q. Wu, and H. Wang. Variance reduction in gradient exploration for online learning to rank. In *Proceedings of the 42nd ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

Z. Wen, B. Kveton, and A. Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

M. Zoghi, T. Tunys, L. Li, D. Jose, J. Chen, C. M. Chin, and M. de Rijke. Click-based hot fixes for underperforming torso queries. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 195–204, 2016.

M. Zoghi, T. Tunys, M. Ghavamzadeh, B. Kveton, C. Szepesvari, and Z. Wen. Online learning to rank in stochastic click models. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

S. Zong, H. Ni, K. Sung, N. R. Ke, Z. Wen, and B. Kveton. Cascading bandits for large-scale recommendation problems. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 2016.

J. Zuo, X. Liu, C. Joe-Wong, J. Lui, and W. Chen. Online competitive influence maximization. *CoRR*, abs/2006.13411, 2020. URL https://arxiv.org/abs/2006.13411.

A TECHNICAL LEMMAS

All technical lemmas and omitted proofs are presented below.

Lemma 4. *Let $X \sim \mathrm{Beta}(\alpha, \beta)$. Then $X - \mathbb{E}[X]$ is σ^2-sub-Gaussian with variance proxy*

$$\sigma^2 = \frac{1}{4(\alpha + \beta + 1)}\,.$$

Proof. This claim is proved in Theorem 1 of Marchal and Arbel (2017). □

Lemma 5. *For any integer n and $a \geq 0$,*

$$\sum_{i=1}^{n} \frac{1}{i + a} \leq \log(1 + n/a)\,.$$

Proof. Since $1/(i + a)$ decreases in i, the sum can be bounded using integration as

$$\sum_{i=1}^{n} \frac{1}{i + a} \leq \int_{x=a}^{n+a} \frac{1}{x} \, \mathrm{d}x = \log(n + a) - \log a = \log(1 + n/a)\,.$$

□

Lemma 6. *For any integer n and $a \geq 0$,*

$$\sum_{i=1}^{n} \frac{1}{\sqrt{i + a}} \leq 2(\sqrt{n + a} - \sqrt{a})\,.$$

Proof. Since $1/\sqrt{i + a}$ decreases in i, the sum can be bounded using integration as

$$\sum_{i=1}^{n} \frac{1}{\sqrt{i + a}} \leq \int_{x=a}^{n+a} \frac{1}{\sqrt{x}} \, \mathrm{d}x = 2(\sqrt{n + a} - \sqrt{a})\,.$$

□

Lemma 7. *Let $(a_i)_{i=1}^K \in [0, 1]^K$ and $(b_i)_{i=1}^K \in [0, 1]^K$. Then*

$$\prod_{i=1}^{K} a_i - \prod_{i=1}^{K} b_i = \sum_{i=1}^{K} \left(\prod_{j=1}^{i-1} a_j\right) (a_i - b_i) \left(\prod_{j=i+1}^{K} b_j\right) = \sum_{i=1}^{K} \left(\prod_{j=1}^{i-1} b_j\right) (a_i - b_i) \left(\prod_{j=i+1}^{K} a_j\right)\,.$$

Proof. The first claim follows from

$$\begin{aligned}\prod_{i=1}^{K} a_i - \prod_{i=1}^{K} b_i &= \prod_{i=1}^{K} a_i - a_1 \prod_{i=2}^{K} b_i + a_1 \prod_{i=2}^{K} b_i - \prod_{i=1}^{K} b_i = a_1 \left(\prod_{i=2}^{K} a_i - \prod_{i=2}^{K} b_i\right) + (a_1 - b_1) \prod_{i=2}^{K} b_i \\ &= \sum_{i=1}^{K} \left(\prod_{j=1}^{i-1} a_j\right) (a_i - b_i) \left(\prod_{j=i+1}^{K} b_j\right)\,,\end{aligned}$$

where the last step is from a recursive application of the same argument to $\left(\prod_{i=2}^K a_i - \prod_{i=2}^K b_i\right)$.

The second claim is an alternative derivation based on

$$\begin{aligned}\prod_{i=1}^{K} a_i - \prod_{i=1}^{K} b_i &= \prod_{i=1}^{K} a_i - b_1 \prod_{i=2}^{K} a_i + b_1 \prod_{i=2}^{K} a_i - \prod_{i=1}^{K} b_i = b_1 \left(\prod_{i=2}^{K} a_i - \prod_{i=2}^{K} b_i\right) + (a_1 - b_1) \prod_{i=2}^{K} a_i \\ &= \sum_{i=1}^{K} \left(\prod_{j=1}^{i-1} b_j\right) (a_i - b_i) \left(\prod_{j=i+1}^{K} a_j\right)\,,\end{aligned}$$

where the last step is from a recursive application of the same argument to $\left(\prod_{i=2}^K a_i - \prod_{i=2}^K b_i\right)$. This concludes the proof. □

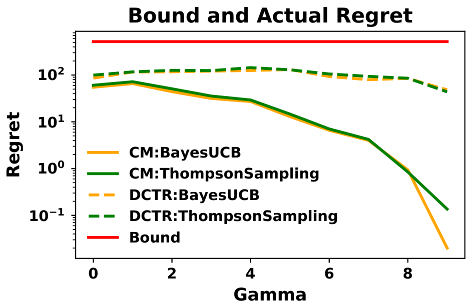

Figure 5: Actual regret and its upper bound as a function of increasing attraction probabilities.

B ADDITIONAL EXPERIMENTS

We include additional results on the effect of the prior, where we vary attraction probabilities. In Figure 5, we show the regret of TS and BayesUCB together with our regret bound, in both the DCTR and CM. The attraction probabilities are sampled as $\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$, where $\alpha_i = 1 + \gamma$ and $\beta_i = 10 - \gamma$. We experiment with $\gamma \in [0, 9]$, where higher γ correspond to problems with higher attraction probabilities. All other parameters are set as in Section 7.2. We observe that our regret bound is pretty tight in the DCTR for all γ. The bound becomes looser in the CM when γ increases, indicating that the regret bound in Theorem 2 can be improved.

C COMPUTATIONAL RESOURCES

The real-world experiment used 200 hours of NVIDIA P100 GPUs for training the offline models and generating the scores for test queries. The bandit algorithm evaluation on test queries took approximately 36k CPU hours. About 97% of that time was spent on evaluating CascadeKL-UCB, which was very slow.

The synthetic experiments took a few minutes on a single CPU.