A Thompson Sampling Algorithm for Cascading Bandits

Wang Chi Cheung
National University of Singapore

Vincent Y. F. Tan
National University of Singapore

Zixin Zhong
National University of Singapore

Abstract

We design and analyze TS-Cascade, a Thompson sampling algorithm for the cascading bandit problem. In TS-Cascade, Bayesian estimates of the click probability are constructed using a univariate Gaussian; this leads to a more efficient exploration procedure vis-à-vis existing UCB-based approaches. We also incorporate the empirical variance of each item's click probability into the Bayesian updates. These two novel features allow us to prove an expected regret bound of the form $\tilde{O}(\sqrt{KLT})$ where L and K are the number of ground items and the number of items in the chosen list respectively and $T \geq L$ is the number of Thompson sampling update steps. This matches the state-of-the-art regret bounds for UCB-based algorithms. More importantly, it is the first theoretical guarantee on a Thompson sampling algorithm for any stochastic combinatorial bandit problem model with partial feedback. Empirical experiments demonstrate superiority of TS-Cascade compared to existing UCB-based procedures in terms of the expected cumulative regret and the time complexity.

1 Introduction

Online recommender systems seek to recommend a small list of items (such as movies or hotels) to users based on a larger ground set $[L] := \{1, \dots, L\}$ of items. The model we consider in this paper is the *cascading bandits* model (Kveton et al., 2015a). In the standard *cascade* model of Craswell et al. (2008), which is used widely in information retrieval and online advertising, the user, upon seeing this list of items, scans through it in a sequential manner. She looks at the first item and if she is *attracted* by it, *clicks* on it. If not, she skips to the next item and clicks on it if she finds it attractive. This process stops when she clicks on one item in the list or when she comes to the end of the list, in which case she is *not attracted* by *any* of the items. The items that are in the ground set but not in the chosen list and those in the list that come after the attractive one are *unobserved.* Each item $i \in [L]$, which has a certain *click probability* $w(i) \in [0, 1]$, attracts the user independently of other items. Under this assumption, the optimal solution is the list of items that maximizes the probability that the user finds an attractive item. This is precisely the list of the most attractive items.

In the *multi-armed bandits* version of the cascade model (Kveton et al., 2015a), the click probabilities $\boldsymbol{w} := \{w(i)\}_{i=1}^{L}$ are *unknown* to the learning agent, and should be learned over time. If the user clicks on any item in the list, a reward of one is obtained by the learning agent. Otherwise, no reward is obtained. Based on the lists previously chosen and the rewards obtained thus far, the agent tries to learn the click probabilities (exploration) in order to adaptively and judiciously recommend other lists of items (exploitation) to maximize his overall reward over T time steps.

Main Contributions. We design and analyze TS-Cascade, a Thompson sampling algorithm (Thompson, 1933) for the cascading bandits problem. Our design involves the two novel features. First, the Bayesian estimates on the vector of latent click probabilities $\boldsymbol{w}$ are constructed by a univariate Gaussian distribution. Consequently, in each time step, Ts-Cascade conducts exploration in a suitably defined one-dimensional space. This leads to a more efficient exploration procedure than the existing Upper Confidence Bound (UCB) approaches, which conduct exploration in L-dimensional confidence hypercubes. Second, inspired by Audibert et al. (2009), we judiciously incorporate the empirical variance of each item's click probability in the Bayesian update. The allows efficient exploration on item i when $w(i)$ is close to 0 or 1.

We establish a problem independent regret bound for our proposed algorithm TS-Cascade. Our regret bound matches the state-of-the-art regret bound for

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

UCB algorithms on the cascading bandit model (Wang and Chen, 2017), up to a multiplicative logarithmic factor in the number of time steps T, when $T \geq L$. Our regret bound is the first theoretical guarantee on a Thompson sampling algorithm for the cascading bandit problem model, or for any stochastic combinatorial bandit problem model with partial feedback (see literature review). Our consideration of Gaussian Thompson sampling is primarily motivated by Zong et al. (2016), who reported the empirical effectiveness of Gaussian Thompson sampling on cascading bandits, and raised its theoretical analysis as an open question. In this paper, we answer this open question in the special case in which there is no linear generalization, and overcome numerous analytical challenges. We carefully design estimates on the latent mean reward (see (4.2)) to handle the subtle statistical dependencies between partial monitoring and Thompson sampling. We reconcile the statistical inconsistency in using Gaussians to model click probabilities by considering a certain truncated version of the Thompson samples (Lemma 4.4). Our framework provides useful tools for analyzing Thompson sampling on stochastic combinatorial bandits with partial feedback in other settings.

Literature Review. Our work is closely related to existing works on the class of stochastic combinatorial bandit (SCB) problems and Thompson sampling. In an SCB model, an arm corresponds to a subset of a ground set of items, each associated with a latent random variable. The corresponding reward depends on the constituent items' realized random variables. SCB models with *semi-bandit feedback*, where a learning agent observes all random variables of the items in a pulled arm, are extensively studied in existing works. Assuming semi-bandit feedback, Anantharam et al. (1987) study the case when the arms constitute a uniform matroid, Kveton et al. (2014) study the case of general matroids, Gai et al. (2010) study the case of permutations, and Gai et al. (2012), Chen et al. (2013), Combes et al. (2015), and Kveton et al. (2015b) investigate various general SCB problem settings. More general settings with contextual information (Li et al. (2010); Qin et al. (2014)) and linear generalization (Wen et al. (2015)) are also studied. All of the works above hinge on UCBs.

Motivated by numerous applications in recommender systems and online advertisement, SCB models have been studied under a more challenging setting of *partial feedback*, where a learning agent only observes the random variables for a subset the items in the pulled arm. A prime example of SCB model with partial feedback is the cascading bandit model, which is first introduced by Kveton et al. (2015a). Subsequently, Kveton et al. (2015c), Katariya et al. (2016), Lagrée et al. (2016) and Zoghi et al. (2017) study the cascading bandit model in various general settings. Cascading bandits with contextual information (Li et al. (2016)) and linear generalization (Zong et al. (2016)) are also studied. Wang and Chen (2017) provide a general algorithmic framework on SCB models with partial feedback. All of the works listed above are also based on UCB.

On the one hand, UCB has been extensively applied for solving various SCB problems. On the other hand, *Thompson sampling* (Thompson, 1933; Chapelle and Li, 2011; Russo et al., 2018), an online algorithm based on Bayesian updates, has been shown to be empirically superior compared to UCB and ϵ-greedy algorithms in various bandit models. The empirical success has motivated a series of research works on the theoretical performance guarantees of Thompson sampling on multi-armed bandits (Agrawal and Goyal, 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2013a, 2017), linear bandits (Agrawal and Goyal, 2013b), generalized linear bandits (Abeille and Lazaric, 2017), etc. Thompson sampling has also been studied for SCB problems with semi-bandit feedback. Komiyama et al. (2015) study the case when the combinatorial arms constitute a uniform matroid; Wang and Chen (2018) investigate the case of general matroids, and Gopalan et al. (2014) and Hüyük and Tekin (2018) consider settings with general reward functions. In addition, SCB problems with semi-bandit feedback are also studied in the Bayesian setting (Russo and Van Roy, 2014), where the latent model parameters are assumed to be drawn from a known prior distribution. Despite existing works, an analysis of Thompson sampling for an SCB problem in the more challenging case of partial feedback is yet to be done. Our work fills in this gap in the literature, and our analysis provides tools for handling the statistical dependence between Thompson sampling and partial feedback in the cascading bandit models.

2 Problem Setup

Let there be $L \in \mathbb{N}$ ground items, denoted as $[L] := \{1, \ldots, L\}$. Each item $i \in [L]$ is associated with a weight $w(i) \in [0, 1]$, signifying the item's click probability. At each time step $t \in [T]$, the agent selects a list of $K \leq L$ items $S_t := (i_1^t, \ldots, i_K^t) \in \pi_K(L)$ to the user, where $\pi_K(L)$ denotes the set of all K-permutations of $[L]$. The user examines the items from i_1^t to i_K^t by examining each item one at a time until possibly all items are examined. For $1 \leq k \leq K$, $W_t(i_k^t) \sim \mathrm{Bern}\left(w(i_k^t)\right)$ are i.i.d. and $W_t(i_k^t) = 1$ iff user clicks on i_k^t at time t.

The *instantaneous reward* of the agent at time t is

$$R(S_t|\boldsymbol{w}) := 1 - \prod_{k=1}^{K} (1 - W_t(i_k^t)) \in \{0, 1\}.$$

In other words, the agent gets a reward of $R(S_t|\boldsymbol{w}) = 1$ if $W_t(i_k^t) = 1$ for some $1 \leq k \leq K$, and a reward of

Table 1: Upper bounds on the T-regret of TS-Cascade, CUCB, CascadeUCB1 and CascadeKL-UCB and the lower bound of all Cascading bandits algorithms.

Algorithm	Reference	Bounds	Problem Indep.
TS-Cascade	Present paper	$O(\sqrt{KLT}\log T + L\log^{5/2} T)$	$\surd$
CUCB	Wang and Chen (2017)	$O\big(\sqrt{KLT}\log T\big)$	$\surd$
CascadeUCB1	Kveton et al. (2015a)	$O\big((L-K)(\log T)/\Delta\big)$	$\times$
CascadeKL-UCB	Kveton et al. (2015a)	$O\big((L-K)\log(T/\Delta)/\Delta\big)$	$\times$
Cascading Bandits	Kveton et al. (2015a)	$\Omega\big((L-K)(\log T)/\Delta\big)$ (Lower Bd)	$\times$

$R(S_t|\boldsymbol{w}) = 0$ if $W_t(i_k^t) = 0$ for all $1 \le k \le K$.

The *feedback* of the agent at time t is defined as

$$k_t := \min\{1 \le k \le K : W_t(i_k^t) = 1\},$$

where we assume that the minimum over an empty set is ∞. If $k_t < \infty$, then the agent observes $W_t(i_k^t) = 0$ for $1 \le k < k_t$, and also observes $W_t(i_k^t) = 1$, but does not observe $W_t(i_k^t)$ for $k > k_t$; otherwise, $k_t = \infty$, then the agent observes $W_t(i_k^t) = 0$ for $1 \le k \le K$.

As the agent aims to maximize the sum of rewards over all steps, a expected cumulative regret is defined to evaluate the performance of an algorithm. First, the *expected instant reward* is

$$r(S|\boldsymbol{w}) = \mathbb{E}[R(S|\boldsymbol{w})] = 1 - \prod_{i_k \in S}(1 - w(i_k)).$$

Note that the expected reward is permutation invariant, but the randomness in the set of observed items is not. Without loss of generality, we assume that $w(1) \ge w(2) \ge \ldots \ge w(L)$, then any permutation of $\{1, \ldots, K\}$ maximizes the mean reward. We let $S^* = (1, \ldots, K)$ be an optimal ordered K-subset for maximizing the expected reward; items in S^* as optimal items and others as suboptimal items. In T steps, we aim to minimize the *expected cumulative regret*:

$$\mathrm{Reg(T)} := T \cdot r(S^*|\boldsymbol{w}) - \sum_{t=1}^{T} r(S_t|\boldsymbol{w}),$$

while the vector of click probabilities $\boldsymbol{w} \in [0,1]^L$ is not known to the agent, and S_t is chosen online, i.e., dependent on previous choices and the previous rewards.

3 Algorithm

Our algorithm is presented in Algorithm 1. Intuitively, to minimize the expected cumulative regret, the agent aims to learn the true weight $w(i)$ of each item $i \in [L]$ by exploring the space to identify S^* (i.e., exploitation) after a hopefully small number of steps. In our algorithm, we approximate the true weight $w(i)$ of each item i by a Bayesian statistic $\theta_t(i)$ at each time step t. This statistic is known as the *Thompson sample.* To do so, first, we sample a one-dimensional standard Gaussian $Z_t \sim \mathcal{N}(0,1)$, define the empirical variance $\hat{\nu}_t(i) = \hat{\mu}_t(i)(1 - \hat{\mu}_t(i))$ of the previously observed arms, and calculate $\theta_t(i)$. Secondly, we select $S_t = (i_1^t, i_2^t, \ldots, i_K^t)$ such that $\theta_t(i_1^t) \ge \theta_t(i_2^t) \ge \cdots \ge \theta_t(i_K^t) \ge \max_{j \notin S_t} \theta_t(j)$; this is reflected in Line 10 of Algorithm 1. Finally, we update the parameters for each observed item i in a standard manner by applying Bayes rule on the mean of the Gaussian (with conjugate prior being another Gaussian) in Line 13.

The algorithm results in the following theoretical guarantee. The proof is sketched in Section 4.

Theorem 3.1. *Consider the cascading bandit problem. Algorithm* TS-Cascade*, presented in Algorithm 1, incurs an expected regret at most*

$$O(\sqrt{KLT}\log T + L\log^{5/2} T),$$

where the big O notation hides a constant factor that is independent of $K, L, T, \boldsymbol{w}$.

In practical applications, $T \gg L$ and so the regret bound is essentially $\tilde{O}(\sqrt{KLT})$. We elaborate on the main features of the algorithm and the guarantee.

In a nutshell, TS-Cascade is a Thompson sampling Algorithm (Thompson, 1933), based on prior-posterior updates on Gaussian random variables with refined variances. The use of the Gaussians is useful, since it allows us to readily generalize the algorithm and analyses to the contextual setting (Li et al., 2010). This handles heterogeneity in the online setting (Li et al., 2016), as well as the linear bandits setting (Zong et al., 2016) for handling a large L. We plan to study these extensions in a future work. To this end, we remark that the posterior update of TS can be done in a variety of ways. While the use of a Beta-Bernoulli update to maintain a Bayesian estimate on $w(i)$ is a natural option (Russo et al., 2018), we use Gaussians instead, in view of their use in generalizations and its empirical success in the linear bandits setting (Zong et al., 2016). Indeed, the conjugate prior-posterior update is not the only choice for TS algorithms for complex multi-armed bandit problems. For example, the posterior update in

Algorithm 1 TS-Cascade, Thompson Sampling for Cascading Bandits with Gaussian Update

```
1: Initialize $\hat{\mu}_1(i) = 0$, $N_1(i) = 0$ for all $i \in [L]$.
2: for $t = 1, 2, \ldots$ do
3:     Sample a 1-dim r.v. $Z_t \sim \mathcal{N}(0, 1)$.
4:     for $i \in [L]$ do
5:         Calculate the empirical variance
               $\hat{\nu}_t(i) = \hat{\mu}_t(i)(1 - \hat{\mu}_t(i)).$
6:         Calculate std. dev. of the Thompson sample
               $\sigma_t(i) = \max\left\{ \sqrt{\frac{\hat{\nu}_t(i)\log(t+1)}{N_t(i)+1}}, \frac{\log(t+1)}{N_t(i)+1} \right\}.$
7:         Construct the Thompson sample
               $\theta_t(i) = \hat{\mu}_t(i) + Z_t \sigma_t(i).$
8:     end for
9:     for $k \in [K]$ do
10:        Extract $i_k^t \in \mathrm{argmax}_{i \in [L] \setminus \{i_1^t, \ldots i_{k-1}^t\}} \theta_t(i)$.
11:    end for
12:    Pull arm $S_t = (i_1^t, i_2^t, \ldots, i_K^t)$.
13:    For each $i \in [L]$, if $W_t(i)$ is observed, define
           $\hat{\mu}_{t+1}(i) = \frac{N_t(i)\hat{\mu}_t(i) + W_t(i)}{N_t(i)+1},$
           $N_{t+1}(i) = N_t(i) + 1.$
       Otherwise, $\hat{\mu}_{t+1}(i) = \hat{\mu}_t(i), N_{t+1}(i) = N_t(i)$.
14: end for
```

Algorithm 2 in Agrawal et al. (2017) for the multinomial lgoit bandit problem is not conjugate.

While the use of Gaussians is useful for generalizations, the analysis of Gaussian Thompson samples in the cascading setting comes with some difficulties, as $\theta_t(i)$ is not in $[0, 1]$ with probability one. We perform a truncation of the Gaussian Thompson sample in the proof of Lemma 4.4 to show that this replacement of the Beta by the Gaussian does not incur any significant loss in terms of the regret and the analysis is not affected significantly.

We elaborate on the refined variances of our Bayesian estimates. Lines 5–7 indicate that the Thompson sample $\theta_t(i)$ is constructed to be a Gaussian random variable with mean $\hat{\mu}_t(i)$ and variance being the maximum of $\hat{\nu}_t(i)\log(t+1)/(N_t(i)+1)$ and $[\log(t+1)/(N_t(i)+1)]^2$. Note that $\hat{\nu}_t(i)$ is the variance of a Bernoulli distribution with mean $\hat{\mu}_t(i)$. In Thompson sampling algorithms, the choice of the variance is of crucial importance. We considered a naïve TS implementation initially. However, the theoretical and empirical results were unsatisfactory, due in part to the large variance of the Thompson sample variance; this motivated us to improve on the algorithm leading to Algorithm 1. The reason why we choose the variance in this manner is to (i) make the Bayesian estimates behave like Bernoulli random variables and to (ii) ensure that it is tuned so that the regret bound has a dependence on $\sqrt{K}$ (see Lemma 4.3) and does not depend on any pre-set parameters. We utilize a key result by Audibert et al. (2009) concerning the analysis of using the empirical variance in multi-arm bandit problems to achieve (i). In essence, in Lemma 4.3, the Thompson sample is shown to depend only on a *single* source of randomness, i.e., the Gaussian random variable Z_t (Line 3 of Algorithm 1). This shaves of a factor of $\sqrt{K}$ vis-à-vis a more naïve analysis where the variance is pre-set in the relevant probability in Lemma 4.3 depends on K independent random variables.

Finally, in Table 1, we compare our regret bound for cascading bandits to those in the literature which are all based on the UCB idea (Wang and Chen, 2017; Kveton et al., 2015a). Note that the last column indicates whether or not the algorithm is problem dependent; being problem dependent means that the bound depends on the vector of click probabilities $\boldsymbol{w}$. To present our results succinctly, for the problem dependent bounds, we assume that the optimal items have the same click probability w_1 and the suboptimal items also have the same click probability $w_2 < w_1$; note though that TS-Cascade makes no such assumption. The gap $\Delta := w_1 - w_2$ is a measure of the difficulty of the problem. Table 1 implies that our upper bound grows like $\sqrt{T}$ just like the others. Our bound also matches the state-of-the-art UCB bound (up to log factors) by Wang and Chen (2017), whose algorithm, when suitably specialized to the cascading bandits setting, is the same as CascadeUCB1 in Kveton et al. (2015a). For the case in which $T \geq L$, our bound is a $\sqrt{\log T}$ factor than the problem independent bound in Wang and Chen (2017), but we are the first to analyze Thompson sampling for the cascading bandits problem.

4 Proof Sketch of Theorem 3.1

In this section, we prove a proof sketch of Theorem 3.1. We also provide the proofs of Lemmas 4.3 and 4.5. The remaining lemmas are proved in Appendix B.

During the iterations, we update $\hat{\mu}_{t+1}(i)$ such that it approaches $w(i)$ eventually. To do so, we select a set S_t according to the order of $\theta_t(i)$'s at each time step. Hence, if $\hat{\mu}_{t+1}(i)$, $\theta_t(i)$ and $w(i)$ are close enough, then we are likely to select the optimal set. This motivates us to define two "nice events" as follows:

$$\mathcal{E}_{\hat{\mu},t} := \{\forall i \in [L] \ : \ |\hat{\mu}_t(i) - w(i)| \leq g_t(i)\},$$
$$\mathcal{E}_{\theta,t} := \{\forall i \in [L] \ : \ |\theta_t(i) - \hat{\mu}_t(i)| \leq h_t(i)\},$$

where $\hat{\nu}_t(i)$ is defined in Line 5 of Algorithm 1, and

$$g_t(i) := \sqrt{\frac{16\hat{\nu}_t(i)\log(t+1)}{N_t(i)+1}} + \frac{24\log(t+1)}{N_t(i)+1},$$
$$h_t(i) := \sqrt{\log(t+1)}g_t(i).$$

Lemma 4.1. *For each $t \in [T]$, $H_t \in \mathcal{E}_{\hat{\mu},t}$, we have*

$$\Pr\left[\mathcal{E}_{\hat{\mu},t}\right] \geq 1 - \frac{3L}{(t+1)^3}, \quad \Pr\left[\mathcal{E}_{\theta,t}|H_t\right] \geq 1 - \frac{1}{2(t+1)^2}.$$

Demonstrating that $\mathcal{E}_{\hat{\mu},t}$ has high probability requires the concentration inequality in Theorem A.1; this is a specialization of a result in Audibert et al. (2009) to Bernoulli random variables. Demonstrating that $\mathcal{E}_{\hat{\theta},t}$ has high probability requires the concentration property of Gaussian random variables (cf. Theorem A.2).

To start our analysis, define

$$F(S,t) := \sum_{k=1}^{K}\left[\prod_{j=1}^{k-1}(1-w(i_j))\right](g_t(i_k)+h_t(i_k)). \quad (4.1)$$

Define the set

$$\mathcal{S}_t := \Big\{S = (i_1,\ldots,i_K) \in \pi_K(L) : \\ F(S,t) \geq r(S^*|\boldsymbol{w}) - r(S|\boldsymbol{w})\Big\}.$$

Recall that $w(1) \geq w(2) \geq \ldots \geq w(L)$. As such, $\mathcal{S}_t$ is non-empty, since $S^* = (1,2,\ldots,K) \in \mathcal{S}_t$.

Intuition behind the set $\mathcal{S}_t$: Ideally, we expect the user to click an item in $\mathcal{S}_t$ for every time step t. Recall that $g_t(i)$ and $h_t(i)$ are decreasing in $N_t(i)$, the number of time steps q's in $1,\ldots,t-1$ when we get to *observe* $W_q(i)$. Naively, arms in $\mathcal{S}_t$ can be thought of as arms that "lack observations", while arms in $\bar{\mathcal{S}}_t$ can be thought of as arms that are "observed enough", and are believed to be suboptimal. Note that $S^* \in \mathcal{S}_t$ is a prime example of an arm that is under-observed.

To further elaborate, $g_t(i) + h_t(i)$ is the "statistical gap" between the Thompson sample $\theta_t(i)$ and the latent mean $w(i)$. The gap shrinks with more observations of i. To balance exploration and exploitation, for any suboptimal item $i \in [L]\setminus[K]$ and any optimal item $k \in [K]$, we should have $g_t(i) + h_t(i) \geq w(k) - w(i)$. However, this is too much to hope for, and it seems that hoping for $S_t \in \mathcal{S}_t$ to happen would be more viable. (See the forthcoming Lemma 4.2.)

Further notations. In addition to set $\mathcal{S}_t$, we define H_t as the collection of observations of the agent, from the beginning until the end of time $t-1$. More precisely, we define $H_t := \{S_q\}_{q=1}^{t-1} \cup \{(i_k^q, W_q(i_k^q))_{k=1}^{\min\{k_t,\infty\}}\}_{q=1}^{t-1}$. Recall that $S_q \in \pi_K(L)$ is the arm pulled during time step q, and $(i_k^q, W_q(i_k^q))_{k=1}^{\min\{k_t,\infty\}}$ is the collection of observed items and their respective values during time step q. At the start of time step t, the agent has observed everything in H_t, and determine the arm S_t to pull accordingly (see Algorithm 1). Note that event $\mathcal{E}_{\hat{\mu},t}$ is $\sigma(H_t)$-measurable. For the convenience of discussion, we define $\mathcal{H}_{\hat{\mu},t} := \{H_t : \text{Event } \mathcal{E}_{\hat{\mu},t} \text{ is true in } H_t\}$. The first statement in Lemma 4.1 is thus $\Pr[H_t \in \mathcal{H}_{\hat{\mu},t}] \geq 1 - 3L/(t+1)^3$.

The performance of Algorithm 1 is analyzed using the following four Lemmas. To begin with, Lemma 4.2 quantifies a set of conditions on $\hat{\boldsymbol{\mu}}_t$ and $\boldsymbol{\theta}_t$ so that the pulled arm S_t belongs to $\mathcal{S}_t$, the collection of arms that lack observations and should be explored. We recall from Lemma 4.1 that the events $\mathcal{E}_{\hat{\mu},t}$ and $\mathcal{E}_{\theta,t}$ hold with high probability. Subsequently, we will crucially use our definition of the Thompson sample $\boldsymbol{\theta}_t$ to argue that inequality (4.2) holds with non-vanishing probability when t is sufficiently large.

Lemma 4.2. *Consider a time step t. Suppose that events $\mathcal{E}_{\hat{\mu},t}, \mathcal{E}_{\theta,t}$ and inequality*

$$\sum_{k=1}^{K}\left[\prod_{j=1}^{k-1}(1-w(j))\right]\theta_t(k) \geq \sum_{k=1}^{K}\left[\prod_{j=1}^{k-1}(1-w(j))\right]w(k) \quad (4.2)$$

hold, then the event $\{S_t \in \mathcal{S}_t\}$ also holds.

In the following, we condition on H_t and show that $\boldsymbol{\theta}_t$ is "typical" w.r.t. $\boldsymbol{w}$ in the sense of (4.2). Due to the conditioning on H_t, the only source of randomness of the pulled arm S_t is from the Thompson sample. Thus, by analyzing a suitably weighted version of the Thompson samples in (4.2), we disentangle the statistical dependence between partial monitoring and Thompson sampling. Recall that $\boldsymbol{\theta}_t$ is normal with $\sigma(H_t)$-measurable mean and variance (Lines 5–7 in Algorithm 1).

Lemma 4.3. *There exists an absolute constant $c \in (0,1)$ independent of $\boldsymbol{w}, K, L, T$ such that, for any time step t and any historical observation $H_t \in \mathcal{H}_{\hat{\mu},t}$, the following inequality holds:*

$$\Pr_{\boldsymbol{\theta}_t}\left[\mathcal{E}_{\theta,t} \text{ and } (4.2) \text{ hold} \mid H_t\right] \geq c - \frac{1}{2(t+1)^3}.$$

Proof. We prove the Lemma by setting the absolute constant c to be $1/(4\sqrt{\pi}e^{8064}) > 0$.

For brevity, we define $\alpha(1) := 1$, and $\alpha(k) = \prod_{j=1}^{k-1}(1-w(j))$ for $2 \leq k \leq K$. By the second part of Lemma 4.1, we know that $\Pr[\mathcal{E}_{\theta,t}|H_t] \geq 1 - 1/2(t+1)^3$, so to complete this proof, it suffices to show that

$\Pr[(4.2) \text{ holds}|H_t] \geq c$. For this purpose, consider

$$\Pr_{\boldsymbol{\theta}_t}\left[\sum_{k=1}^{K}\alpha(k)\theta_t(k) \geq \sum_{k=1}^{K}\alpha(k)w(k) \;\middle|\; H_t\right]$$

$$= \Pr_{Z_t}\left[\sum_{k=1}^{K}\alpha(k)\left[\hat{\mu}_t(k)+Z_t\sigma_t(k)\right] \geq \sum_{k=1}^{K}\alpha(k)w(k) \;\middle|\; H_t\right] \tag{4.3}$$

$$\geq \Pr_{Z_t}\left[\sum_{k=1}^{K}\alpha(k)\left[w(k)-g_t(k)\right] + Z_t\cdot\sum_{k=1}^{K}\alpha(k)\sigma_t(k) \geq \sum_{k=1}^{K}\alpha(k)w(k) \;\middle|\; H_t\right] \tag{4.4}$$

$$= \Pr_{Z_t}\left[Z_t\cdot\sum_{k=1}^{K}\alpha(k)\sigma_t(k) \geq \sum_{k=1}^{K}\alpha(k)g_t(k) \;\middle|\; H_t\right]$$

$$\geq \frac{1}{4\sqrt{\pi}}\exp\left\{-\frac{7}{2}\left[\sum_{k=1}^{K}\alpha(k)g_t(k)\Big/\sum_{k=1}^{K}\alpha(k)\sigma_t(k)\right]^2\right\} \tag{4.5}$$

$$\geq \frac{1}{4\sqrt{\pi}}\exp\left\{-\frac{7}{2}\left[\sum_{k=1}^{K}\alpha(k)g_t(k)\Big/ \sum_{k=1}^{K}\alpha(k)\left(\frac{1}{2}\sqrt{\frac{\hat{\nu}_t(i)\log(t+1)}{N_t(i)+1}}+\frac{1}{2}\frac{\log(t+1)}{(N_t(i)+1)}\right)\right]^2\right\}$$

$$\geq \frac{1}{4\sqrt{\pi}e^{8064}} = c. \tag{4.6}$$

Step (4.3) is by the definition of $\{\theta_t(i)\}_{i\in L}$ in Line 7 in Algorithm 1. It is important to note that these samples share the *same* random seed Z_t. Next, step (4.4) is by the Lemma assumption that $H_t \in \mathcal{H}_{\hat{\mu},t}$, which means that $\hat{\mu}_t(k) \geq w_t(k) - g_t(k)$ for all $k \in [K]$. Step (4.5) is an application of the anti-concentration inequality of a normal random variable in Theorem A.2. Step (4.6) is by applying the definition of $g_t(i)$. □

Combining Lemmas 4.2 and 4.3, we conclude that there exists an absolute constant c such that, for any time step t and any historical observation $H_t \in \mathcal{H}_{\hat{\mu},t}$,

$$\Pr_{\boldsymbol{\theta}_t}\left[S_t \in \mathcal{S}_t \;\middle|\; H_t\right] \geq c - \frac{1}{2(t+1)^3}. \tag{4.7}$$

Equipped with (4.7), we are able to provide an upper bound on the regret of our Thompson sampling algorithm at every sufficiently large time step.

Lemma 4.4. *Let c be an absolute constant such that Lemma 4.3 holds true. Consider a time step t that satisfies $c-1/(t+1)^3 > 0$. Conditional on an arbitrary but fixed historical observation $H_t \in \mathcal{H}_{\hat{\mu},t}$, we have*

$$\mathbb{E}_{\boldsymbol{\theta}_t}[r(S^*|\boldsymbol{w}) - r(S_t|\boldsymbol{w})|H_t] \leq \left(1+\frac{4}{c}\right)\mathbb{E}_{\boldsymbol{\theta}_t}\left[F(S_t,t) \;\middle|\; H_t\right] + \frac{L}{2(t+1)^2}.$$

The proof of Lemma 4.4 relies crucially on *truncating* the original Thompson sample $\boldsymbol{\theta}_t \in \mathbb{R}$ to $\tilde{\boldsymbol{\theta}}_t \in [0,1]^L$. Under this truncation operation, S_t remains optimal under $\tilde{\boldsymbol{\theta}}_t$ (as it was under $\boldsymbol{\theta}_t$) and $|\tilde{\theta}_t(i) - w(i)| \leq |\theta_t(i) - w(i)|$, i.e., the distance from the truncated Thompson sample to the ground truth is not increased.

For any t satisfying $c - 1/(t+1)^3 > 0$, define

$$\mathcal{F}_{i,t} := \{\text{Observe } W_t(i) \text{ at } t\}$$

$$G(S_t, \boldsymbol{W}_t) := \sum_{k=1}^{K} 1\big(\mathcal{F}_{i_k^t,t}\big)\cdot\big(g_t(i_k^t)+h_t(i_k^t))\big),$$

we unravel the upper bound in Lemma 4.4 to establish the expected regret at time step t:

$$\begin{aligned}
&\mathbb{E}\left\{r(S^*|\boldsymbol{w}) - r(S_t|\boldsymbol{w})\right\} \\
&\leq \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\theta}_t}[r(S^*|\boldsymbol{w}) - r(S_t|\boldsymbol{w}) \mid H_t]\cdot 1(H_t \in \mathcal{H}_{\hat{\mu},t})\right] \\
&\qquad + \mathbb{E}\left[1(H_t \notin \mathcal{H}_{\hat{\mu},t})\right] \\
&\leq \left(1+\frac{4}{c}\right)\mathbb{E}\left[\mathbb{E}_{\boldsymbol{\theta}_t}\left[F(S_t,t) \mid H_t\right]1(H_t \in \mathcal{H}_{\hat{\mu},t})\right] \\
&\qquad + \frac{1}{2(t+1)^2} + \frac{3L}{(t+1)^3} \\
&\leq \left(1+\frac{4}{c}\right)\mathbb{E}\left[F(S_t,t)\right] + \frac{4L}{(t+1)^2} &(4.8) \\
&= \left(1+\frac{4}{c}\right)\mathbb{E}\left[\mathbb{E}_{\boldsymbol{W}_t}[G(S_t,\boldsymbol{W}_t)\big|H_t,S_t]\right] + \frac{4L}{(t+1)^2} \\
&= \left(1+\frac{4}{c}\right)\mathbb{E}\left[G(S_t,\boldsymbol{W}_t)\right] + \frac{4L}{(t+1)^2}, &(4.9)
\end{aligned}$$

where (4.8) follows by assuming t is sufficiently large.

Lemma 4.5. *For any realization of historical trajectory $\mathcal{H}_{T+1}$, we have*

$$\sum_{t=1}^{T} G(S_t, \boldsymbol{W}_t) \leq 6\sqrt{KLT}\log T + 144L\log^{5/2} T.$$

Proof. Recall that for each $i \in [L]$ and $t \in [T+1]$, $N_t(i) = \sum_{s=1}^{t-1} 1(\mathcal{F}_{i,s})$ is the number of rounds in $[t-1]$ when we get to observe the outcome for item i. Since $G(S_t, \boldsymbol{W}_t)$ involves $g_t(i) + h_t(i)$, we first bound this term. The definitions of $g_t(i)$ and $h_t(i)$ yield that

$$g_t(i) + h_t(i) \leq \frac{12\log(t+1)}{\sqrt{N_t(i)+1}} + \frac{72\log^{3/2}(t+1)}{N_t(i)+1}.$$

Subsequently, we decompose $\sum_{t=1}^{T} G(S_t, \boldsymbol{W}_t)$ according to its definition. For a fixed but arbitrary item i, consider the sequence $(U_t(i))_{t=1}^T = (1(\mathcal{F}_{i,t}) \cdot (g_t(i) + h_t(i)))_{t=1}^T$. Clearly, $U_t(i) \neq 0$ if and only if the decision maker observes the realization $W_t(i)$ of item i at t. Let $t = \tau_1 < \tau_2 < \ldots < \tau_{N_{T+1}}$ be the time steps when $U_t(i) \neq 0$. We assert that $N_{\tau_n}(i) = n-1$ for each n. Indeed, prior to time steps τ_n, item i is observed precisely in the time steps $\tau_1, \ldots, \tau_{n-1}$. Thus, we have

$$\sum_{t=1}^{T} 1(\mathcal{F}_{i,t}) \cdot (g_t(i) + h_t(i)) = \sum_{n=1}^{N_{T+1}(i)} (g_{\tau_n}(i) + h_{\tau_n}(i))$$
$$\leq \sum_{n=1}^{N_{T+1}(i)} \frac{12\log T}{\sqrt{n}} + \frac{72\log^{3/2} T}{n}. \tag{4.10}$$

Now we complete the proof as follows:

$$\sum_{t=1}^{T}\sum_{k=1}^{K} 1(\mathcal{F}_{i_k^t,t}) \cdot (g_t(i_k^t) + h_t(i_k^t))$$
$$= \sum_{i\in[L]}\sum_{t=1}^{T} 1(\mathcal{F}_{i,t}) \cdot (g_t(i) + h_t(i))$$
$$\leq \sum_{i\in[L]}\sum_{n=1}^{N_{T+1}(i)} \frac{12\log T}{\sqrt{n}} + \frac{72\log^{3/2} T}{n} \tag{4.11}$$
$$\leq 6\sum_{i\in[L]}\sqrt{N_{T+1}(i)}\log T + 72L\log^{3/2} T(\log T + 1)$$
$$\leq 6\sqrt{L\sum_{i\in[L]} N_{T+1}(i)\log T + 72L\log^{3/2} T(\log T + 1)} \tag{4.12}$$
$$\leq 6\sqrt{KLT}\log T + 144L\log^{5/2} T, \tag{4.13}$$

where (4.11) follows from (4.10), (4.12) follows from the Cauchy-Schwarz inequality, and (4.13) is because the decision maker can observe at most K items at each time step, hence $\sum_{i\in[L]} N_{T+1}(i) \leq KT$. □

Finally, we bound the total regret from above by considering the time step $t_0 := \lceil 1/c^{1/3} \rceil$, and then bound the regret for the time steps before t_0 by 1 and the regret for time steps after by inequality (4.9), which holds for all $t > t_0$:

$$\mathrm{Reg}(T) \leq \left\lceil \frac{1}{c^{1/3}} \right\rceil + \sum_{t=t_0+1}^{T} \mathbb{E}\left\{r(S^*|\boldsymbol{w}) - r(S_t|\boldsymbol{w})\right\}$$
$$\leq \left\lceil \frac{1}{c^{1/3}} \right\rceil + \left(1 + \frac{4}{c}\right)\mathbb{E}\left[\sum_{t=1}^{T} G(S_t, \boldsymbol{W}_t)\right]$$
$$+ \sum_{t=t_0+1}^{T} \frac{4L}{(t+1)^2}.$$

It is clear that the third term is $O(L)$, and by Lemma 4.5, the second term is $O(\sqrt{KLT}\log T + L\log^{5/2} T)$. Altogether, Theorem 3.1 is proved.

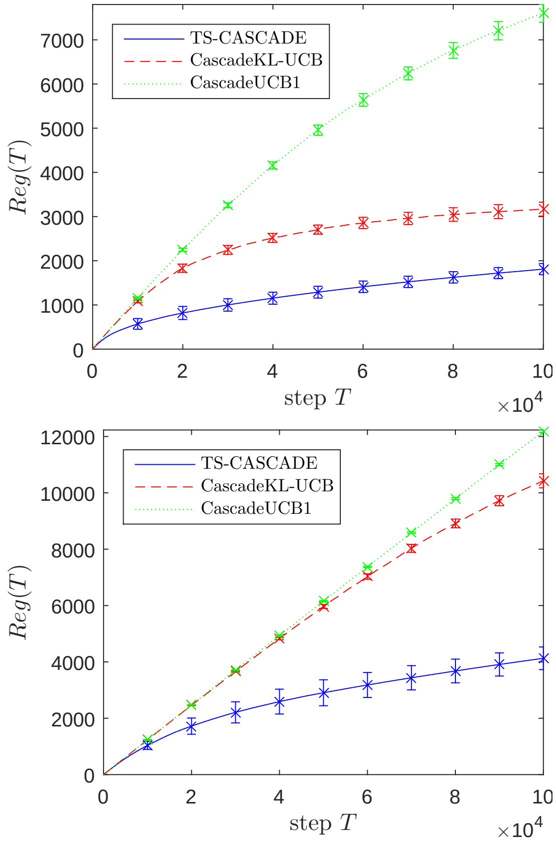

Figure 1: Reg(T) of TS-Cascade, CascadeKL-UCB and CascadeUCB1 with $L \in \{64, 256\}$ (resp. top and bottom), $K = 2$ and $\Delta = 0.075$. Each line indicates the average Reg(T) (over 20 runs) and the length of each errorbar above and below each data point is the standard deviation.

5 Experiments

In this section, we evaluate the performance of TS-Cascade using numerical simulations. To demonstrate the effectiveness of our algorithm, we compare the expected cumulative regret of TS-Cascade to CascadeKL-UCB and CascadeUCB1 in Kveton et al. (2015a). We reimplemented the latter two algorithms and checked that their performances are roughly the same as those in Table 1 of Kveton et al. (2015a).

We set the optimal items to have the same click probability w_1 and the suboptimal items to also have the same click probability $w_2 < w_1$. The gap $\Delta := w_1 - w_2$. We set $w_1 = 0.2$, $T = 10^5$ and vary L, K, and Δ. We conduct 20 independent simulations with each algorithm under each setting of L, K, and Δ. We calculate the average and standard deviation of Reg(T), and as well as the average running time of each experiment. Here we only present a subset of the results. More details are given in Appendix C.

In Table 2, we compare the performances of algorithms

Table 2: The performances of TS-Cascade, CascadeKL-UCB and CascadeUCB1 under 18 different settings. For each algorithm, the first column shows the mean and the standard deviation of $\mathrm{Reg}(T)$ and the second column shows the average running time in seconds. For each problem setting, the algorithm with smallest average $\mathrm{Reg}(T)$ and shortest running time is marked in bold.

L	K	Δ	TS-Cascade		CascadeKL-UCB		CascadeUCB1	
16	2	0.15	377.07 ± 11.67	3.16	**359.35 ± 26.42**	54.3	1277.42 ± 25.88	**2.82**
16	4	0.15	294.55 ± 15.08	3.03	**265.9 ± 20.36**	54.48	990.51 ± 31.72	**2.84**
16	8	0.15	**138.85 ± 9.81**	3.51	148.36 ± 12.35	55.5	555.83 ± 14.41	**3.17**
32	2	0.15	**738.19 ± 19.23**	3.41	764.42 ± 48.57	105.4	2711.44 ± 58.41	**2.98**
32	4	0.15	**612.36 ± 10.66**	3.55	619.68 ± 34.56	105.56	2237.77 ± 43.7	**3.02**
32	8	0.15	**381.8 ± 13.19**	3.68	419.39 ± 19.59	105.64	1526.97 ± 24.48	**3.14**
32	2	0.075	**1159 ± 63.43**	**3.49**	1583.33 ± 104.04	106.62	4217.87 ± 129.08	3.95
32	4	0.075	**1062.9 ± 80.06**	**3.55**	1208.06 ± 59.25	106.08	3301.44 ± 85.43	3.84
32	8	0.075	**631.45 ± 51.51**	**3.58**	718.65 ± 32.27	106.51	1890.06 ± 47.8	3.97
64	2	0.075	**1810.43 ± 126.74**	4.74	3169.17 ± 156.98	207.31	7599.58 ± 199.99	**4.24**
64	4	0.075	**1730.13 ± 128.09**	**4.88**	2512.28 ± 106.85	208.08	6437.43 ± 239.96	5.04
64	8	0.075	**1175.07 ± 46.91**	**4.7**	1565.76 ± 72.98	208.34	3962.35 ± 87.61	4.77
128	2	0.075	**2784.44 ± 185.08**	5.36	6160.86 ± 300.48	414.45	11055.68 ± 156.27	**5.17**
128	4	0.075	**2837.25 ± 239.41**	4.76	5004.45 ± 188.68	412.55	11516.47 ± 227.48	**4.7**
128	8	0.075	**2004.58 ± 122.26**	4.87	3084.67 ± 105.78	413.6	7432.14 ± 129.24	**4.61**
256	2	0.075	**4128.96 ± 400.88**	8.35	10426.63 ± 249.33	816.52	12191.23 ± 39.69	**7.22**
256	4	0.075	**4376.73 ± 373.99**	**7.49**	9389.72 ± 251.5	818.07	15748.08 ± 131.08	7.56
256	8	0.075	**3258.24 ± 238.91**	**7.24**	6019.24 ± 145.95	820	12417.86 ± 160.53	7.83

under 18 different settings. Since CascadeKL-UCB perfoms far better than CascadeUCB1, we mainly focus on the comparison between our method and CascadeKL-UCB. In most cases, the expected cumulative regret of our algorithm is significantly smaller than that of CascadeKL-UCB, especially when L is large and Δ is small. Note that a larger L means that the problem size is larger. A smaller Δ implies that the difference between optimal and sub-optimal arms are less pronounced. Hence, when L is large and Δ is small, the problem is "more difficult". However, the standard deviation of our algorithm is larger than that of CascadeKL-UCB in some cases. A possible explanation is that Thompson sampling yields more randomness than UCB due to the additional randomness of the Thompson samples $\{\boldsymbol{\theta}_t\}_{t\in[T]}$. In contrast, UCB-based algorithms do not have this source of randomness as each upper confidence bound is deterministically designed. Furthermore, Table 2 suggests that our algorithm is much faster than CascadeKL-UCB and is just as fast as CascadeUCB1. The reason why CascadeKL-UCB is so slow is because an UCB has to be computed via an optimization problem for every $i \in [L]$. In contrast, TS-Cascade in Algorithm 1 does not contain any computationally expensive steps.

In Figure 1, we plot $\mathrm{Reg}(T)$ as a function of T for TS-Cascade, CascadeKL-UCB and CascadeUCB1 when $L \in \{64, 256\}$, $K = 2$ and $\Delta = 0.075$. It is clear that our method outperforms the two UCB algorithms. For the case where the number of ground items $L = 256$ is large, the UCB-based algorithms do not demonstrate the $\sqrt{T}$ behavior even after $T = 10^5$ iterations. In contrast, $\mathrm{Reg}(T)$ for TS-Cascade behaves as $O(\sqrt{T})$ which implies that the empirical performance corroborates the upper bound derived in Theorem 3.1. We have plotted $\mathrm{Reg}(T)$ for other settings of L, K and Δ in Appendix C and the same conclusion can be drawn.

6 Summary and Future work

This work presents the first theoretical analysis of Thompson sampling for cascading bandits. The expected regret matches the state-of-the-art based on UCB by Wang and Chen (2017) (which is identical to CascadeUCB1 in Kveton et al. (2015a)). Empirical experiments, however, show the clear superiority of TS-Cascade over CascadeKL-UCB and CascadeUCB1 in terms of regret and running time.

From Table 2, we see that a problem-independent lower bound is still not available. It is envisioned that a judicious construction of an adversarial bandit example, together with the information-theoretic technique of (Auer et al., 2002, Theorem 5.1) will lead to a lower bound of the form $\Omega(\sqrt{KLT})$, matching Theorem 3.1 here and Wang and Chen (2017). Next, we envision that a refinement of the proof techniques herein, especially the design of Thompson samples to be Gaussian, would be useful for generalization the contextual setting (Li et al., 2010; Qin et al., 2014; Li et al., 2016).

References

M. Abeille and A. Lazaric. Linear Thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 176–184, 2017.

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth edition, 1964.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 39.1–39.26, 2012.

S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31, pages 99–107, 2013a.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 127–135, 2013b.

S. Agrawal and N. Goyal. Near-optimal regret bounds for Thompson sampling. *J. ACM*, 64(5):30:1–30:24, Sept. 2017.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977*, 2017.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays–Part I: I.I.D. rewards. *IEEE Transactions on Automatic Control*, 32 (11):968–976, November 1987.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration-–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876 – 1902, 2009. Algorithmic Learning Theory.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77, 2002.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, 2013.

R. Combes, M. S. Talebi, A. Proutière, and M. Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems 28*, 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.

Y. Gai, B. Krishnamachari, and R. Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN)*, pages 1–9, April 2010.

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, Oct 2012.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 100–108, 2014.

A. Hüyük and C. Tekin. Thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms. *Manuscript*, 2018. https://arxiv.org/abs/1809.02707.

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*, volume 48, pages 1215–1224, 2016.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213, 2012.

J. Komiyama, J. Honda, and H. Nakagawa. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of The 32nd International Conference on Machine Learning*, volume 37, pages 1152–1161, 2015.

B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 420–429, 2014.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, 2015a.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 535–543, 2015b.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvári. Combinatorial cascading bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1450–1458, 2015c.

P. Lagrée, C. Vernade, and O. Cappe. Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems 29*, pages 1597–1605. Curran Associates, Inc., 2016.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

S. Li, B. Wang, S. Zhang, and W. Chen. Contextual combinatorial cascading bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1245–1253, 2016.

L. Qin, S. Chen, and X. Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *SDM*, pages 461–469. SIAM, 2014.

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

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

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.

Q. Wang and W. Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pages 1161–1171, 2017.

S. Wang and W. Chen. Thompson sampling for combinatorial semi-bandits. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5114–5122, 2018.

Z. Wen, B. Kveton, and A. Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1113–1122, 2015.

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*, volume 70, pages 4199–4208, 06–11 Aug 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 Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 835–844, 2016.