Analysis of Thompson Sampling for Combinatorial Multi-armed Bandit with Probabilistically Triggered Arms

Alihan Hüyük
Electrical and Electronics
Engineering Department
Bilkent University, Ankara, Turkey

Cem Tekin
Electrical and Electronics
Engineering Department
Bilkent University, Ankara, Turkey

Abstract

We analyze the regret of combinatorial Thompson sampling (CTS) for the combinatorial multi-armed bandit with probabilistically triggered arms under the semi-bandit feedback setting. We assume that the learner has access to an exact optimization oracle but does not know the expected base arm outcomes beforehand. When the expected reward function is Lipschitz continuous in the expected base arm outcomes, we derive $O(\sum_{i=1}^{m} \log T/(p_i \Delta_i))$ regret bound for CTS, where m denotes the number of base arms, p_i denotes the minimum non-zero triggering probability of base arm i and Δ_i denotes the minimum suboptimality gap of base arm i. We also compare CTS with combinatorial upper confidence bound (CUCB) via numerical experiments on a cascading bandit problem.

1 INTRODUCTION

Multi-armed bandit (MAB) exhibits the prime example of the tradeoff between exploration and exploitation faced in many reinforcement learning problems [1]. In the classical MAB, at each round the learner selects an arm which yields a random reward that comes from an unknown distribution. The goal of the learner is to maximize its expected cumulative reward over all rounds by learning to select arms that yield high rewards. The learner's performance is measured by its regret with respect to an oracle policy which always selects the arm with the highest expected reward. It is shown that when the arms' rewards are independent,

any uniformly good policy will incur at least logarithmic in time regret [2]. Several classes of policies are proposed for the learner to minimize its regret. These include Thompson sampling [3, 4, 5] and upper confidence bound (UCB) policies [2, 6, 7], which are shown to achieve logarithmic in time regret, and hence, are order optimal.

Combinatorial multi-armed bandit (CMAB) [8, 9, 10] is an extension of MAB where the learner selects a *super arm* at each round, which is defined to be a subset of the *base arms*. Then, the learner observes and collects the reward associated with the selected super arm, and also observes the outcomes of the base arms that are in the selected super arm. This type of feedback is also called the *semi-bandit* feedback. For the special case when the expected reward of a super arm is a linear combination of the expected outcomes of the base arms that are in that super arm, it is shown in [9] that a combinatorial version of UCB1 in [7] achieves $O(Km \log T/\Delta)$ gap-dependent and $O(\sqrt{KmT \log T})$ gap-free regrets, where m is the number of base arms, K is the maximum number of base arms in a super arm, and Δ is the gap between the expected reward of the optimal super arm and the second best super arm.

Later on, this setting is generalized to allow the expected reward of the super arm to be a more general function of the expected outcomes of the base arms that obeys certain monotonicity and bounded smoothness conditions [10]. The main challenge in the general case is that the optimization problem itself is NP-hard, but an approximately optimal solution can usually be computed efficiently for many special cases [11]. Therefore, it is assumed that the learner has access to an approximation oracle, which can output a super arm that has expected reward that is at least α fraction of the optimal reward with probability at least β when given the expected outcomes of the base arms. Thus, the regret is measured with respect to the $\alpha\beta$ fraction of the optimal reward, and it is proven that a combinatorial variant of UCB1, called CUCB, achieves

$O(\sum_{i=1}^{m} \log T / \Delta_i)$ regret, when the bounded smoothness function is $f(x) = \gamma x$ for some $\gamma > 0$, where Δ_i is the minimum gap between the expected reward of the optimal super arm and the expected reward of any suboptimal super arm that contains base arm i. Recently, it is shown in [12] that Thompson sampling can achieve $O(\sum_{i=1}^{m} \log T / \Delta_i)$ regret for the general CMAB under a Lipschitz continuity assumption on the expected reward, given that the learner has access to an exact computation oracle, which outputs an optimal super arm when given the set of expected base arm outcomes. Moreover, it is also shown that the learner cannot guarantee sublinear regret when it only has access to an approximation oracle.

An interesting extension of CMAB is CMAB with probabilistically triggered arms (PTAs) [13] where the selected super arm probabilistically triggers a set of base arms, and the reward obtained in a round is a function of the set of probabilistically triggered base arms and their expected outcomes. For this problem, it is shown in [13] that logarithmic regret is achievable when the expected reward function has the l_∞ bounded smoothness property. However, this bound depends on $1/p^*$, where p^* is the minimum non-zero triggering probability. Later, it is shown in [14] that under a more strict smoothness assumption on the expected reward function, called triggering probability modulated bounded smoothness, it is possible to achieve regret which does not depend on $1/p^*$. It is also shown in this work that the dependence on $1/p^*$ is unavoidable for the general case. In another work [15], CMAB with PTAs is considered for the case when the arm triggering probabilities are all positive, and it is shown that both CUCB and CTS achieve bounded regret.

Apart from the works mentioned above, numerous other works also tackle related online learning problems. For instance, [16] considers matroid MAB, which is a special case of CMAB where the super arms are given as independent sets of a matroid with base arms being the elements of the ground set, and the expected reward of a super arm is the sum of the expected outcomes of the base arms in the super arm. In addition, Thompson sampling is also analyzed for a parametric CMAB model given a prior with finite support in [17], and a contextual CMAB model with a Bayesian regret metric in [18]. Unlike these works, we adopt the models in [12] and [13], work in a setting where there is an unknown but fixed parameter (expected outcome) vector, and analyze the expected regret.

To sum up, in this work we analyze the (expected) regret of CTS when the learner has access to an exact computation oracle, and prove that it achieves $O(\sum_{i=1}^{m} \log T / (p_i \Delta_i))$ regret. Comparing this with the regret lower bound for CMAB with PTAs given in The-

orem 3 in [14], we also observe that our regret bound is tight.

The rest of the paper is organized as follows. Problem formulation is given in Section 2. CTS algorithm is described in Section 3. Regret analysis of CTS is given in Section 4. Numerical results are given in Section 5, and concluding remarks are given in Section 6. Additional numerical results and proofs of the lemmas that are used in the regret analysis are given in the supplemental document.

2 PROBLEM FORMULATION

CMAB is a decision making problem where the learner interacts with its environment through m base arms, indexed by the set $[m] := \{1, 2, ..., m\}$ sequentially over rounds indexed by $t \in [T]$. In this paper, we consider the general CMAB model introduced in [13] and borrow the notation from [12]. In this model, the following events take place in order in each round t:

- The learner selects a subset of base arms, denoted by $S(t)$, which is called a super arm.

- $S(t)$ causes some other base arms to probabilistically trigger based on a stochastic triggering process, which results in a set of triggered base arms $S'(t)$ that contains $S(t)$.

- The learner obtains a reward that depends on $S'(t)$ and observes the outcomes of the arms in $S'(t)$.

Next, we describe in detail the arm outcomes, the super arms, the triggering process, and the reward.

At each round t, the environment draws a random outcome vector $\boldsymbol{X}(t) := (X_1(t), X_2(t), \ldots, X_m(t))$ from a fixed probability distribution D on $[0, 1]^m$ independent of the previous rounds, where $X_i(t)$ represents the outcome of base arm i. D is unknown by the learner, but it belongs to a class of distributions $\mathcal{D}$ which is known by the learner. We define the mean outcome (parameter) vector as $\boldsymbol{\mu} := (\mu_1, \mu_2, \ldots, \mu_m)$, where $\mu_i := \mathbb{E}_{\boldsymbol{X} \sim D}[X_i(t)]$, and use $\boldsymbol{\mu}_S$ to denote the projection of $\boldsymbol{\mu}$ on S for $S \subseteq [m]$.

Since CTS computes a posterior over $\boldsymbol{\mu}$, the following assumption is made to have an efficient and simple update of the posterior distribution.

Assumption 1. *The outcomes of all base arms are mutually independent, i.e., $D = D_1 \times D_2 \times \cdots \times D_m$.*

Note that this independence assumption is correct for many applications, including the influence maximization problem with independent cascade influence propagation model [19].

The learner is allowed to select $S(t)$ from a subset of $2^{[m]}$ denoted by $\mathcal{I}$, which corresponds to the set of feasible super arms. Once $S(t)$ is selected, all base arms $i \in S(t)$ are immediately triggered. These arms can trigger other base arms that are not in $S(t)$, and those arms can further trigger other base arms, and so on. At the end, a random superset $S'(t)$ of $S(t)$ is formed that consists of all triggered base arms as a result of selecting $S(t)$. We have $S'(t) \sim D^{\mathrm{trig}}(S(t), \boldsymbol{X}(t))$, where D^{trig} is the probabilistic triggering function that describes the triggering process. For instance, in the influence maximization problem, D^{trig} may correspond to the independent cascade influence propagation model defined over a given influence graph [19]. The triggering process can also be described by a set of triggering probabilities. Essentially, for each $i \in [m]$ and $S \in \mathcal{I}$, $p_i^{D',S}$ denotes the probability that base arm i is triggered when super arm S is selected given that the arm outcome distribution is $D' \in \mathcal{D}$. For simplicity, we let $p_i^S = p_i^{D,S}$, where D is the true arm outcome distribution. Let $\tilde{S} := \{i \in [m] : p_i^S > 0\}$ be the set of all base arms that could potentially be triggered by super arm S, which is called the *triggering set* of S. We have that $S(t) \subseteq S'(t) \subseteq \tilde{S}(t) \subseteq [m]$. We define $p_i := \min_{S \in \mathcal{I}: i \in \tilde{S}} p_i^S$ as the minimum nonzero triggering probability of base arm i, and $p^* := \min_{i \in [m]} p_i$ as the minimum nonzero triggering probability.

At the end of round t, the learner receives a reward that depends on the set of triggered arms $S'(t)$ and the outcome vector $\boldsymbol{X}(t)$, which is denoted by $R(S'(t), \boldsymbol{X}(t))$. For simplicity of notation, we also use $R(t) = R(S'(t), \boldsymbol{X}(t))$ to denote the reward in round t. Note that whether a base arm is in the selected super arm or is triggered afterwards is not relevant in terms of the reward. We also make two other assumptions about the reward function R, which are standard in the CMAB literature [12, 13].

Assumption 2. *The expected reward of super arm $S \in \mathcal{I}$ only depends on S and the mean outcome vector $\boldsymbol{\mu}$, i.e., there exists a function r such that*

$$\mathbb{E}[R(t)] = \mathbb{E}_{S'(t) \sim D^{trig}(S(t), \boldsymbol{X}(t)), \boldsymbol{X}(t) \sim D}[R(S'(t), \boldsymbol{X}(t))]$$
$$= r(S(t), \boldsymbol{\mu}).$$

Assumption 3. *(Lipschitz continuity) There exists a constant $B > 0$, such that for every super arm S and every pair of mean outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, $|r(S, \boldsymbol{\mu}) - r(S, \boldsymbol{\mu}')| \le B \| \boldsymbol{\mu}_{\tilde{S}} - \boldsymbol{\mu}'_{\tilde{S}} \|_1$, where $\| \cdot \|_1$ denotes the l_1 norm.*

We consider the semi-bandit feedback model, where at the end of round t, the learner observes the individual outcomes of the triggered arms, denoted by $Q(S'(t), \boldsymbol{X}(t)) := \{(i, X_i(t)) : i \in S'(t)\}$. Again, for simplicity of notation, we also use $Q(t) =$

$Q(S'(t), \boldsymbol{X}(t))$ to denote the observation at the end of round t. Based on this, the only information available to the learner when choosing the super arm to select in round $t + 1$ is its observation history, given as $\mathcal{F}_t := \{(S(\tau), Q(\tau)) : \tau \in [t]\}$.

In short, the tuple $([m], \mathcal{I}, D, D^{\mathrm{trig}}, R)$ constitutes a CMAB-PTA problem instance. Among the elements of this tuple only D is unknown to the learner.

In order to evaluate the performance of the learner, we define the set of optimal super arms given an m-dimensional parameter vector $\boldsymbol{\theta}$ as $\mathrm{OPT}(\boldsymbol{\theta}) := \mathrm{argmax}_{S \in \mathcal{I}} r(S, \boldsymbol{\theta})$. We use $\mathrm{OPT} := \mathrm{OPT}(\boldsymbol{\mu})$ to denote the set of optimal super arms given the true mean outcome vector $\boldsymbol{\mu}$. Based on this, we let S^* to represent a specific super arm in $\mathrm{argmin}_{S \in \mathrm{OPT}} |\tilde{S}|$, which is the set of super arms that have triggering sets with minimum cardinality among all optimal super arms. We also let $k^* := |S^*|$ and $\tilde{k}^* := |\tilde{S}^*|$.

Next, we define the suboptimality gap due to selecting super arm $S \in \mathcal{I}$ as $\Delta_S := r(S^*, \mu) - r(S, \mu)$, the maximum suboptimality gap as $\Delta_{\max} := \max_{S \in \mathcal{I}} \Delta_S$, and the minimum suboptimality gap of base arm i as $\Delta_i := \min_{S \in \mathcal{I} - \mathrm{OPT}: i \in \tilde{S}} \Delta_S$.[1] The goal of the learner is to minimize the (expected) regret over the time horizon T, given by

$$\mathrm{Reg}(T) := \mathbb{E}\left[\sum_{t=1}^{T} (r(S^*, \boldsymbol{\mu}) - r(S(t), \boldsymbol{\mu}))\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{S(t)}\right]. \tag{1}$$

3 THE LEARNING ALGORITHM

We consider the CTS algorithm for CMAB with PTAs [12, 15] (pseudocode given in Algorithm 1). We assume that the learner has access to an exact computation oracle, which takes as input a parameter vector $\boldsymbol{\theta}$ and the problem structure $([m], \mathcal{I}, D^{\mathrm{trig}}, R)$, and outputs a super arm, denoted by $\mathrm{Oracle}(\boldsymbol{\theta})$, such that $\mathrm{Oracle}(\boldsymbol{\theta}) \in \mathrm{OPT}(\boldsymbol{\theta})$. CTS keeps a Beta posterior over the mean outcome of each base arm. At the beginning of round t, for each base arm i it draws a sample $\theta_i(t)$ from its posterior distribution. Then, it forms the parameter vector in round t as $\boldsymbol{\theta}(t) := (\theta_1(t), \ldots, \theta_m(t))$, gives it to the exact computational oracle, and selects the super arm $S(t) = \mathrm{Oracle}(\boldsymbol{\theta}(t))$. At the end of the round, CTS updates the posterior distributions of the triggered base arms using the observation $Q(t)$.

---

[1] If there is no such super arm S, let $\Delta_i = \infty$.

Algorithm 1 Combinatorial Thompson Sampling (CTS).

1: For each base arm i, let $a_i = 1$, $b_i = 1$
2: **for** $t = 1, 2, \ldots$ **do**
3:    For each base arm i, draw a sample $\theta_i(t)$ from Beta distribution $\beta(a_i, b_i)$; let $\boldsymbol{\theta}(t) := (\theta_1(t), \ldots, \theta_m(t))$
4:    Select super arm $S(t) = \text{Oracle}(\boldsymbol{\theta}(t))$, get the observation $Q(t)$
5:    **for all** $(i, X_i) \in Q(t)$ **do**
6:       $Y_i \leftarrow 1$ with probability X_i, 0 with probability $1 - X_i$
7:       $a_i \leftarrow a_i + Y_i$
8:       $b_i \leftarrow b_i + (1 - Y_i)$
9:    **end for**
10: **end for**

4 REGRET ANALYSIS

4.1 Main Theorem

The regret bound for CTS is given in the following theorem.

Theorem 1. *Under Assumptions 1, 2 and 3, for all D, the regret of CTS by round T is bounded as follows:*

$$Reg(T) \leq \sum_{i=1}^{m} \max_{S \in \mathcal{I} - OPT : i \in \tilde{S}} \frac{16B^2 |\tilde{S}| \log T}{(1 - \rho)p_i(\Delta_S - 2B(\tilde{k}^{*2} + 2)\varepsilon)}$$

$$+ \left(3 + \frac{\tilde{K}^2}{(1 - \rho)p^* \varepsilon^2} + \frac{2\mathbb{I}\{p^* < 1\}}{\rho^2 p^*} \right) m \Delta_{\max}$$

$$+ \alpha \frac{8\tilde{k}^*}{p^* \varepsilon^2} \left(\frac{4}{\varepsilon^2} + 1 \right)^{\tilde{k}^*} \log \frac{\tilde{k}^*}{\varepsilon^2} \Delta_{\max}$$

for all $\rho \in (0, 1)$, and for all $0 < \varepsilon \leq 1/\sqrt{e}$ such that $\forall S \in \mathcal{I} - OPT$, $\Delta_S > 2B(\tilde{k}^{*2} + 2)\varepsilon$, where B is the Lipschitz constant in Assumption 3, $\alpha > 0$ is a problem independent constant that is also independent of T, and $\tilde{K} := \max_{S \in \mathcal{I}} |\tilde{S}|$ is the maximum triggering set size among all super arms.

We compare the result of Theorem 1 with [13], which shows that the regret of CUCB is $O(\sum_{i \in [m]} \log T / (p_i \Delta_i))$, given an l_∞ bounded smoothness condition on the expected reward function, when the bounded smoothness function is $f(x) = \gamma x$. When ε is sufficiently small, the regret bound in Theorem 1 is asymptotically equivalent to the regret bound for CUCB (in terms of the dependence on T, p_i and Δ_i, $i \in [m]$).

For the case with $p^* = 1$ (no probabilistic triggering), the regret bound in Theorem 1 matches with the regret bound in Theorem 1 in [12] (in terms of the dependence on T and Δ_i, $i \in [m]$). As a final remark, we note that

Theorem 3 in [14] shows that the $1/p_i$ term in the regret bound that multiplies the $\log T$ term is unavoidable in general.

4.2 Preliminaries for the Proof

The complement of set $\mathcal{S}$ is denoted by $\neg \mathcal{S}$ or $\mathcal{S}^c$. The indicator function is given as $\mathbb{I}\{\cdot\}$. $M_i(t) := \sum_{\tau=1}^{t-1} \mathbb{I}\{i \in \tilde{S}(\tau)\}$ denotes the number of times base arm i is in the triggering set of the selected super arm, i.e., it is tried to be triggered, $N_i(t) := \sum_{\tau=1}^{t-1} \mathbb{I}\{i \in S'(\tau)\}$ denotes the number of times base arm i is triggered, and $\hat{\mu}_i(t) := \frac{1}{N_i(t)} \sum_{\tau : \tau < t, i \in S'(\tau)} Y_i(\tau)$ denotes the empirical mean outcome of base arm i until round t, where $Y_i(t)$ is the Bernoulli random variable with mean $X_i(t)$ that is used for updating the posterior distribution that corresponds to base arm i in CTS.

We define

$$\ell(S) := \frac{2 \log T}{\left(\frac{\Delta_S}{2B|\tilde{S}|} - \frac{\tilde{k}^{*2} + 2}{|\tilde{S}|} \varepsilon \right)^2}$$

as the *sampling threshold* of super arm S,

$$L_i(S) := \frac{\ell(S)}{(1 - \rho)p_i} \tag{2}$$

as the *trial threshold* of base arm i with respect to super arm S, and $L_i^{\max} := \max_{S \in \mathcal{I} - OPT : i \in \tilde{S}} L_i(S)$.

Consider an m-dimensional parameter vector $\boldsymbol{\theta}$. Similar to [12], given $Z \subseteq \tilde{S}^*$, we say that the *first bad event for Z*, denoted by $\mathcal{E}_{Z,1}(\boldsymbol{\theta})$, holds when all $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_Z, \boldsymbol{\theta}_{Z^c})$ such that $\|\boldsymbol{\theta}'_Z - \boldsymbol{\mu}_Z\|_\infty \leq \varepsilon$ satisfies the following properties:

- $Z \subseteq \tilde{\text{Oracle}}(\boldsymbol{\theta}')$.

- Either $\text{Oracle}(\boldsymbol{\theta}') \in OPT$ or $\|\boldsymbol{\theta}'_{\tilde{\text{Oracle}}(\boldsymbol{\theta}')} - \boldsymbol{\mu}_{\tilde{\text{Oracle}}(\boldsymbol{\theta}')}\|_1 > \frac{\Delta_{\tilde{\text{Oracle}}(\boldsymbol{\theta}')}}{B} - (\tilde{k}^{*2} + 1)\varepsilon$.

Given the same parameter vector $\boldsymbol{\theta}$, the *second bad event for Z* is defined as $\mathcal{E}_{Z,2}(\boldsymbol{\theta}) := \|\boldsymbol{\theta}_Z - \boldsymbol{\mu}_Z\|_\infty > \varepsilon$.

In addition, similar to the regret analysis in [12], we will make use of the following events when bounding the regret:

$$\mathcal{A}(t) := \{S(t) \notin OPT\} \tag{3}$$

$$\mathcal{B}_{i,1}(t) := \left\{ |\hat{\mu}_i(t) - \mu_i| > \frac{\varepsilon}{|\tilde{S}(t)|} \right\}$$

$$\mathcal{B}_{i,2}(t) := \{N_i(t) \leq (1 - \rho)p_i M_i(t)\} \tag{4}$$

$$\mathcal{B}(t) := \left\{ \exists i \in \tilde{S}(t) : \mathcal{B}_{i,1}(t) \vee \mathcal{B}_{i,2}(t) \right\} \tag{5}$$

$$\mathcal{C}(t) := \left\{ \exists i \in \tilde{S}(t) : |\theta_i(t) - \hat{\mu}_i(t)| > \sqrt{\frac{2 \log T}{N_i(t)}} \right\} \tag{6}$$

$$\mathcal{D}(t) := \left\{ \|\boldsymbol{\theta}_{\tilde{S}(t)}(t) - \boldsymbol{\mu}_{\tilde{S}(t)}\|_1 > \frac{\Delta_{S(t)}}{B} - (\tilde{k}^{*2} + 1)\varepsilon \right\}$$
$$(7)$$

4.3 Regret Decomposition

Using the definitions of the events given in (3)-(7), the regret can be upper bounded as follows:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{A}(t)\} \times \Delta_{S(t)}]$$

$$\leq \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{B}(t) \wedge \mathcal{A}(t)\} \times \Delta_{S(t)}] \qquad (8)$$

$$+ \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{C}(t) \wedge \mathcal{A}(t)\} \times \Delta_{S(t)}] \qquad (9)$$

$$+ \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\neg\mathcal{B}(t) \wedge \neg\mathcal{C}(t) \wedge \mathcal{D}(t) \wedge \mathcal{A}(t)\}$$

$$\times \Delta_{S(t)}] \qquad (10)$$

$$+ \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\neg\mathcal{D}(t) \wedge \mathcal{A}(t)\} \times \Delta_{S(t)}] . \qquad (11)$$

The regret bound in Theorem 1 is obtained by bounding the terms in the above decomposition. All events except the one specified in (10) can only happen a small (finite) number of times. Every time (10) happens, there must be base arms in the triggering set of the selected super arm which are "tried" to be triggered less than the trial threshold. These "under-explored" base arms are the main contributors of the regret, and their contribution depends on how many times they are "tried". Moreover, every time these base arms are tried, their contribution to the future regret decreases. Thus, by summing up these contributions we obtain a logarithmic bound for (10). In the proof, we will make use of the facts and lemmas that are introduced in the following section.

4.4 Facts and Lemmas

Fact 1. *(Lemma 4 in [12]) When CTS is run, the following holds for all base arms $i \in [m]$:*

$$\Pr\left[\theta_i(t) - \hat{\mu}_i(t) > \sqrt{\frac{2\log T}{N_i(t)}}\right] \leq \frac{1}{T}$$

$$\Pr\left[\hat{\mu}_i(t) - \theta_i(t) > \sqrt{\frac{2\log T}{N_i(t)}}\right] \leq \frac{1}{T}$$

We also have the following three lemmas (proofs can be found in the supplemental document along with some additional facts that are used in the proofs).

Lemma 1. *When CTS is run, we have*

$$\mathbb{E}[|\{t : 1 \leq t \leq T, i \in \tilde{S}(t), |\hat{\mu}_i(t) - \mu_i| > \varepsilon \vee \mathcal{B}_{i,2}(t)\}|]$$

$$\leq 1 + \frac{1}{(1-\rho)p^*\varepsilon^2} + \frac{2\mathbb{I}\{p^* < 1\}}{\rho^2 p^*}$$

for all $i \in [m]$ and $\rho \in (0, 1)$.

Lemma 2. *Suppose that $\neg\mathcal{D}(t) \wedge \mathcal{A}(t)$ happens. Then, there exists $Z \subseteq \tilde{S}^*$ such that $Z \neq \emptyset$ and $\mathcal{E}_{Z,1}(\boldsymbol{\theta}(t))$ holds.*

Lemma 3. *When CTS is run, for all $Z \subseteq \tilde{S}^*$ such that $Z \neq \emptyset$, we have*

$$\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{E}_{Z,1}(\boldsymbol{\theta}(t)), \mathcal{E}_{Z,2}(\boldsymbol{\theta}(t))\}]$$

$$\leq 13\alpha_2' \cdot \frac{|Z|}{p^*} \cdot \left(\frac{2^{2|Z|+3} \log \frac{|Z|}{\varepsilon^2}}{\varepsilon^{2|Z|+2}} \right)$$

where α_2' is a problem independent constant.

4.5 Main Proof of Theorem 1

4.5.1 Bounding (8)

Using Lemma 1, we have

$$\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{B}(t) \wedge \mathcal{A}(t)\} \times \Delta_{S(t)}]$$

$$\leq \Delta_{\max} \sum_{i=1}^{m} \mathbb{E}\left[\left|\left\{t : 1 \leq t \leq T, i \in \tilde{S}(t), \right.\right.\right.$$

$$\left.\left.\left. |\hat{\mu}_i(t) - \mu_i| > \frac{\varepsilon}{\tilde{K}} \vee \mathcal{B}_{i,2}(t) \right\}\right|\right]$$

$$\leq \left(1 + \frac{\tilde{K}^2}{(1-\rho)p^*\varepsilon^2} + \frac{2\mathbb{I}\{p^* < 1\}}{\rho^2 p^*} \right) m\Delta_{\max}.$$

4.5.2 Bounding (9)

By Fact 1, we have

$$\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{C}(t) \wedge \mathcal{A}(t)\} \times \Delta_{S(t)}]$$

$$\leq \Delta_{\max} \sum_{i=1}^{m} \sum_{t=1}^{T} \Pr\left[|\theta_i(t) - \hat{\mu}_i(t)| > \sqrt{\frac{2\log T}{N_i(t)}} \right]$$

$$\leq 2m\Delta_{\max}.$$

4.5.3 Bounding (10)

For this, we first show that event $\neg\mathcal{B}(t) \wedge \neg\mathcal{C}(t) \wedge \mathcal{D}(t) \wedge \mathcal{A}(t)$ cannot happen when $M_i(t) > L_i(S(t)), \forall i \in \tilde{S}(t)$. To see this, assume that both $\neg\mathcal{B}(t) \wedge \neg\mathcal{C}(t) \wedge \mathcal{A}(t)$ and $M_i(t) > L_i(S(t)), \forall i \in \tilde{S}(t)$ hold. Then, we must have

$$\|\boldsymbol{\theta}_{\tilde{S}(t)}(t) - \hat{\boldsymbol{\mu}}_{\tilde{S}(t)}(t)\|_1$$

$$= \sum_{i \in \tilde{S}(t)} |\theta_i(t) - \hat{\mu}_i(t)|$$

$$\leq \sum_{i \in \tilde{S}(t)} \sqrt{\frac{2 \log T}{N_i(t)}} \tag{12}$$

$$\leq \sum_{i \in \tilde{S}(t)} \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(t)}} \tag{13}$$

$$\leq \sum_{i \in \tilde{S}(t)} \sqrt{\frac{2 \log T}{(1 - \rho) p_i L_i(S(t))}}$$

$$= \sum_{i \in \tilde{S}(t)} \sqrt{\frac{2 \log T}{\ell(S(t))}} \tag{14}$$

$$= \sum_{i \in \tilde{S}(t)} \left(\frac{\Delta_{S(t)}}{2B|\tilde{S}(t)|} - \frac{\tilde{k}^{*2} + 2}{|\tilde{S}(t)|} \varepsilon \right)$$

$$= |\tilde{S}(t)| \left(\frac{\Delta_{S(t)}}{2B|\tilde{S}(t)|} - \frac{\tilde{k}^{*2} + 2}{|\tilde{S}(t)|} \varepsilon \right)$$

$$= \frac{\Delta_{S(t)}}{2B} - (\tilde{k}^{*2} + 2)\varepsilon$$

where (12) holds when $\neg \mathcal{C}(t)$ happens, (13) holds when $\neg \mathcal{B}(t)$ happens, and (14) holds by the definition of $L_i(S(t))$.

We also know that $\|\hat{\boldsymbol{\mu}}_{\tilde{S}(t)}(t) - \boldsymbol{\mu}_{\tilde{S}(t)}\|_1 \leq \varepsilon$, when $\neg \mathcal{B}(t)$ happens. Then, $\|\boldsymbol{\theta}_{\tilde{S}(t)}(t) - \boldsymbol{\mu}_{\tilde{S}(t)}\|_1 \leq \|\boldsymbol{\theta}_{\tilde{S}(t)}(t) - \hat{\boldsymbol{\mu}}_{\tilde{S}(t)}(t)\|_1 + \|\hat{\boldsymbol{\mu}}_{\tilde{S}(t)}(t) - \boldsymbol{\mu}_{\tilde{S}(t)}\|_1 \leq \frac{\Delta_{S(t)}}{2B} - (\tilde{k}^{*2} + 1)\varepsilon < \frac{\Delta_{S(t)}}{B} - (\tilde{k}^{*2} + 1)\varepsilon$, which implies that $\neg \mathcal{D}(t)$ happens. Thus, we conclude that when $\neg \mathcal{B}(t) \wedge \neg \mathcal{C}(t) \wedge \mathcal{D}(t) \wedge \mathcal{A}(t)$ happens, then there exists some $i \in \tilde{S}(t)$ such that $M_i(t) \leq L_i(S(t))$. Let $S_1(t)$ be the base arms i in $\tilde{S}(t)$ such that $M_i(t) > L_i(S(t))$, and $S_2(t)$ be the other base arms in $\tilde{S}(t)$. By the result above, $S_2(t) \neq \emptyset$. Next, we show that

$$\Delta_{S(t)} \leq 2B \sum_{i \in S_2(t)} \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(t)}}.$$

This holds since,

$$\frac{\Delta_{S(t)}}{B} - (\tilde{k}^{*2} + 1)\varepsilon < \sum_{i \in \tilde{S}(t)} |\theta_i(t) - \mu_i|$$

$$\leq \sum_{i \in \tilde{S}(t)} (|\theta_i(t) - \hat{\mu}_i(t)| + |\hat{\mu}_i(t) - \mu_i|)$$

$$\leq \sum_{i \in S_1(t)} |\theta_i(t) - \hat{\mu}_i(t)| + \sum_{i \in S_2(t)} |\theta_i(t) - \hat{\mu}_i(t)| + \varepsilon$$

$$\leq |S_1(t)| \left(\frac{\Delta_{S(t)}}{2B|\tilde{S}(t)|} - \frac{\tilde{k}^{*2} + 2}{|\tilde{S}(t)|} \varepsilon \right)$$

$$+ \sum_{i \in S_2(t)} |\theta_i(t) - \hat{\mu}_i(t)| + \varepsilon$$

$$\leq \frac{\Delta_{S(t)}}{2B} - (\tilde{k}^{*2} + 1)\varepsilon + \sum_{i \in S_2(t)} |\theta_i(t) - \hat{\mu}_i(t)|$$

$$\leq \frac{\Delta_{S(t)}}{2B} - (\tilde{k}^{*2} + 1)\varepsilon + \sum_{i \in S_2(t)} \sqrt{\frac{2 \log T}{N_i(t)}}$$

$$\leq \frac{\Delta_{S(t)}}{2B} - (\tilde{k}^{*2} + 1)\varepsilon + \sum_{i \in S_2(t)} \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(t)}}.$$

Fix $i \in [m]$. For $w > 0$, let η_w^i be the round for which $i \in S_2(\eta_w^i)$ and $|\{t \leq \eta_w^i : i \in S_2(t)\}| = w$, and $w^i(T) := |\{t \leq T : i \in S_2(t)\}|$. We have $i \in S_2(\eta_w^i) \subseteq \tilde{S}(\eta_w^i)$ for all $w > 0$, which implies that $M_i(\eta_{w+1}^i) \geq w$. Moreover, by the definition of $S_2(t)$, we know that $M_i(t) \leq L_i(S(t)) \leq L_i^{\max}$ for $i \in S_2(t)$, $t \leq T$. These two facts together imply that $w^i(T) \leq L_i^{\max}$ with probability 1.

Consider the round τ_1^i for which $i \in \tilde{S}(t)$ for the first time, i.e., $\tau_1^i := \min\{t : i \in \tilde{S}(t)\}$. We know that $M_i(\tau_1^i) = 0 \leq L_i(S)$ for all S, hence $i \in S_2(\tau_1^i)$. Since $\forall t < \tau_1^i$, $i \notin \tilde{S}(t)$, and $i \notin \tilde{S}(t)$ implies $i \notin S_2(t)$, we conclude that $\tau_1^i = \eta_1^i$. We also observe that $\neg \mathcal{B}(t)$ cannot happen for $t \leq \tau_1^i = \eta_1^i$, since $N_i(t) > (1 - \rho) p_i M_i(t) = 0$ cannot be true when $N_i(t) \leq M_i(t) = 0$.

Then,

$$\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\neg \mathcal{B}(t) \wedge \neg \mathcal{C}(t) \wedge \mathcal{D}(t) \wedge \mathcal{A}(t)\} \times \Delta_{S(t)}]$$

$$\leq \mathbb{E} \left[\sum_{t=1}^{T} \sum_{i \in S_2(t)} \mathbb{I}\{\neg \mathcal{B}(t)\} 2B \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(t)}} \right]$$

$$= \mathbb{E} \left[\sum_{i=1}^{m} \sum_{t=1}^{T} \mathbb{I}\{i \in S_2(t), \neg \mathcal{B}(t)\} \right.$$

$$\left. \times 2B \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(t)}} \right]$$

$$\leq \mathbb{E} \left[\sum_{i=1}^{m} \sum_{t=\eta_1^i + 1}^{T} \mathbb{I}\{i \in S_2(t)\} 2B \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(t)}} \right]$$

$$\leq \mathbb{E} \left[\sum_{i=1}^{m} \sum_{w=1}^{\lfloor L_i^{\max} \rfloor} \sum_{t=\eta_w^i + 1}^{\eta_{w+1}^i} \mathbb{I}\{i \in S_2(t)\} \right.$$

$$\left. \times 2B \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(t)}} \right]$$

$$= \mathbb{E} \left[\sum_{i=1}^{m} \sum_{w=1}^{\lfloor L_i^{\max} \rfloor} 2B \sqrt{\frac{2 \log T}{(1 - \rho) p_i M_i(\eta_{w+1}^i)}} \right]$$

$$\leq \sum_{i=1}^{m} \sum_{w=1}^{\lfloor L_i^{\max} \rfloor} 2B \sqrt{\frac{2 \log T}{(1 - \rho) p_i w}}$$

$$\le \sum_{i=1}^{m} 4B \sqrt{\frac{2 \log T}{(1-\rho)p_i} L_i^{\max}} \qquad (15)$$

where (15) holds since $\sum_{n=1}^{N} \sqrt{1/n} \le 2\sqrt{N}$.

4.5.4 Bounding (11)

From Lemma 2, we know that

$$\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\neg\mathcal{D}(t) \wedge \mathcal{A}(t)\} \times \Delta_{S(t)}]$$

$$\le \Delta_{\max} \sum_{Z \subseteq \tilde{S}^*, Z \neq \emptyset} \left(\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{E}_{Z,1}(\boldsymbol{\theta}(t)), \mathcal{E}_{Z,2}(\boldsymbol{\theta}(t))\}] \right)$$

since $\neg\mathcal{D}(t) \wedge \mathcal{A}(t)$ implies $\mathcal{E}_{Z,1}(\boldsymbol{\theta}(t))$ for some $Z \subseteq \tilde{S}^*$, and $\mathcal{E}_{Z,1}(\boldsymbol{\theta}(t)) \wedge \neg\mathcal{E}_{Z,2}(\boldsymbol{\theta}(t))$ implies either $\neg\mathcal{A}(t)$ or $\mathcal{D}(t)$.

From Lemma 3, we have:

$$\sum_{Z \subseteq \tilde{S}^*, Z \neq \emptyset} \left(\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\mathcal{E}_{Z,1}(\boldsymbol{\theta}(t)), \mathcal{E}_{Z,2}(\boldsymbol{\theta}(t))\}] \right)$$

$$\le \sum_{Z \subseteq \tilde{S}^*, Z \neq \emptyset} 13\alpha_2' \cdot \frac{|Z|}{p^*} \cdot \left(\frac{2^{2|Z|+3} \log \frac{|Z|}{\varepsilon^2}}{\varepsilon^{2|Z|+2}} \right)$$

$$\le 13\alpha_2' \frac{8\tilde{k}^*}{p^*\varepsilon^2} \log \frac{\tilde{k}^*}{\varepsilon^2} \sum_{Z \subseteq \tilde{S}^*, Z \neq \emptyset} \frac{2^{2|Z|}}{\varepsilon^{2|Z|}}$$

$$\le 13\alpha_2' \frac{8\tilde{k}^*}{p^*\varepsilon^2} \left(\frac{4}{\varepsilon^2} + 1 \right)^{\tilde{k}^*} \log \frac{\tilde{k}^*}{\varepsilon^2}.$$

4.5.5 Summing the Bounds

The regret bound for CTS is computed by summing the bounds derived for terms (8)-(11) in the regret decomposition, which are given in the sections above:

$$\mathrm{Reg}(T) \le \sum_{i=1}^{m} 4B \sqrt{\frac{2 \log T}{(1-\rho)p_i} L_i^{\max}}$$

$$+ \left(3 + \frac{\tilde{K}^2}{(1-\rho)p^*\varepsilon^2} + \frac{2\mathbb{I}\{p^* < 1\}}{\rho^2 p^*} \right) m\Delta_{\max}$$

$$+ 13\alpha_2' \frac{8\tilde{k}^*}{p^*\varepsilon^2} \left(\frac{4}{\varepsilon^2} + 1 \right)^{\tilde{k}^*} \log \frac{\tilde{k}^*}{\varepsilon^2} \Delta_{\max}$$

$$= \sum_{i=1}^{m} \max_{S \in \mathcal{I}-\mathrm{OPT}: i \in \tilde{S}} \frac{16B^2 |\tilde{S}| \log T}{(1-\rho)p_i(\Delta_S - 2B(\tilde{k}^{*2} + 2)\varepsilon)}$$

$$+ \left(3 + \frac{\tilde{K}^2}{(1-\rho)p^*\varepsilon^2} + \frac{2\mathbb{I}\{p^* < 1\}}{\rho^2 p^*} \right) m\Delta_{\max}$$

$$+ \alpha \frac{8\tilde{k}^*}{p^*\varepsilon^2} \left(\frac{4}{\varepsilon^2} + 1 \right)^{\tilde{k}^*} \log \frac{\tilde{k}^*}{\varepsilon^2} \Delta_{\max}$$

where $\alpha := 13\alpha_2'$.

4.6 Differences from the Analysis in [12]

Our regret analysis differs from the analysis in [12] (without probabilistic triggering) in the following aspects: First of all, our bad events $\mathcal{E}_{Z,1}(\boldsymbol{\theta})$ and $\mathcal{E}_{Z,2}(\boldsymbol{\theta})$ given in Section 4.2 are defined in terms of subsets Z of $\tilde{S}^*$ rather than S^*. Secondly, we need to relate the number of times base arm i is in the triggering set of the selected super arm $(M_i(t))$ with the number of times it is triggered $(N_i(t))$, which requires us to define events $\mathcal{B}_{i,2}(t)$ for $i \in [m]$ as given in (4), and use them in the regret decomposition. This introduces new challanges in bounding (10), where we make use of a variable called *trial threshold* $(L_i(S(t)))$ given in (2) to show that (10) cannot happen when $M_i(t) > L_i(S(t))$, $\forall i \in \tilde{S}(t)$. We also need to take probabilistic triggering into account when proving Lemmas 1 and 3. For instance, in Lemma 3, we define a new way to count the number of times $\mathcal{E}_{Z,1}(\boldsymbol{\theta}) \wedge \mathcal{E}_{Z,2}(\boldsymbol{\theta})$ happens for all $Z \subseteq \tilde{S}^*$ such that $Z \neq \emptyset$.

5 NUMERICAL RESULTS

In this section, we compare CTS with CUCB in [13] in a cascading bandit problem [20], which is a special case of CMAB with PTAs. In the disjunctive form of this problem a search engine outputs a list of K web pages for each of its L users among a set of R web pages. Then, the users examine their respective lists, and click on the first page that they find attractive. If all pages fail to attract them, they do not click on any page. The goal of the search engine is to maximize the number of clicks.

The problem can be modeled as a CMAB problem. The base arms are user-page pairs (i, j), where $i \in [L]$ and $j \in [R]$. User i finds page j attractive independent of other users and other pages, and the probability that user i finds page j attractive is given as $p_{i,j}$. The super arms are L lists of K-tuples, where each K-tuple represents the list of pages shown to a user. Given a super arm S, let $S(i, k)$ denote the kth page that is selected for user i. Then, the triggering probabilities can be written as

$$p_{(i,j)}^S = \begin{cases} 1 & \text{if } j = S(i,1) \\ \prod_{k'=1}^{k-1}(1 - p_{i,S(i,k')}) & \text{if } \exists k \neq 1 : j = S(i,k) \\ 0 & \text{otherwise} \end{cases}$$

that is we observe feedback for a top selection immediately, and observe feedback for the other selections only if all previous selections fail to attract the user. The expected reward of playing super arm S can be written as $r(S, \boldsymbol{p}) = \sum_{i=1}^{L} \left(1 - \prod_{k=1}^{K}(1 - p_{i,S(i,k)}) \right)$ for which Assumption 3 holds when $B = 1$.

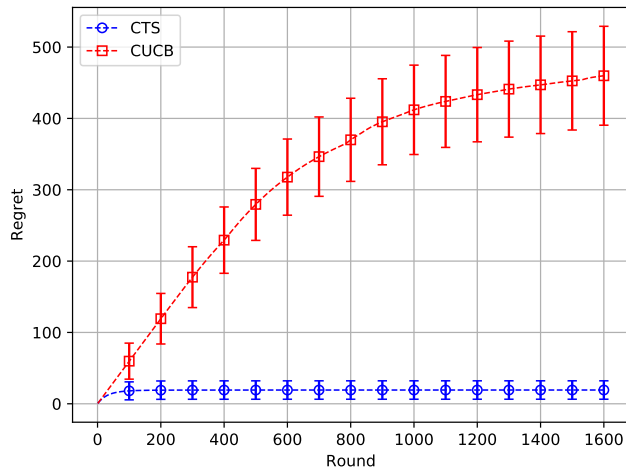Figure 1: Regrets of CTS and CUCB for the disjunctive cascading bandit problem.

Figure 2: Regrets of CTS and CUCB for the conjunctive cascading bandit problem.

We let $L = 20$, $R = 100$ and $K = 5$, and generate $p_{i,j}$s by sampling uniformly at random from $[0, 1]$. We run both CTS and CUCB for 1600 rounds, and report their regrets averaged over 1000 runs in Figure 1, where error bars represent the standard deviation of the regret (multiplied by 10 for visibility). As expected CTS significantly outperforms CUCB. Relatively bad performance of CUCB can be explained by excessive number of explorations due to the UCBs that stay high for large number of rounds.

Next, we study the conjunctive analogue of the problem that we consider, where the goal of the search engine is to maximize the number of users with lists that do not contain unattractive pages, and when examining their lists, users provide feedback by reporting the first unattractive page. Formally,

$$p_{(i,j)}^S = \begin{cases} 1 & \text{if } j = S(i,1) \\ \prod_{k'=1}^{k-1} p_{i,S(i,k')} & \text{if } \exists k \neq 1 : j = S(i,k) \\ 0 & \text{otherwise} \end{cases}$$

and $r(S, \boldsymbol{p}) = \sum_{i=1}^{L} \prod_{k=1}^{K} p_{i,S(i,k)}$. In this setting, we let $L = 1$, $R = 1000000$, $K = 999999$, and set all $p_{i,j} = 1$ except for one, which is set to $1/3$. Essentially, we are trying to find the one page that does not certainly attract the user. We run both CTS and CUCB for 160 rounds, and report their regrets averaged over 1000 runs in Figure 2, where error bars represent the standard deviation of the regret. It is observed that CUCB outperforms CTS in the first 80 rounds and CTS outperforms CUCB after the first 80 rounds.

This case is specifically constructed to show that CTS does not always uniformly outperform CUCB when there are many good arms but a single bad arm. For 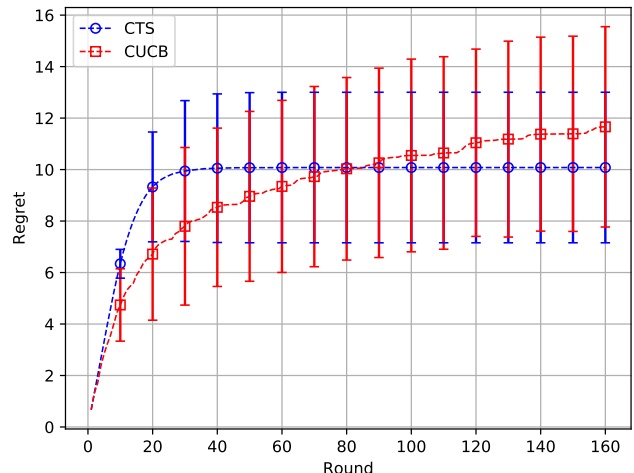CTS to act optimally in this case, the sample from the posterior of the bad arm should be less than the samples from the posteriors of the other arms, which happens with a small probability at the beginning. CUCB is good at the beginning because of the "deterministic" indices. Whenever the bad arm is observed, with high probability its UCB falls below the UCBs of the good arms, and hence, the bad arm is "forgotten" for some time, which leads to better initial performance.

6 CONCLUSION

In this paper we analyzed the regret of CTS for CMAB with PTAs. We proved an order optimal regret bound when the expected reward function is Lipschitz continuous without assuming monotonicity of the expected reward function. Our bound includes the $1/p^*$ term that is unavoidable in general. Future work includes deriving regret bounds under more strict assumptions on the expected reward function such as the triggering probability modulated bounded smoothness condition given in [14] to get rid of the $1/p^*$ term. This is especially challenging due to the fact that the triggering probabilities depend on the expected arm outcomes, which makes relating the expected rewards under different realizations of $\boldsymbol{\theta}$ (such as given in the proof of Lemma 2) difficult.

References

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[2] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[3] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[4] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. 25th Annual Conference on Learning Theory*, pp. 39.1–39.26, 2012.

[5] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[6] R. Agrawal, "Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

[8] 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*, vol. 20, no. 5, pp. 1466–1478, 2012.

[9] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proc. 18th International Conference on Artificial Intelligence and Statistics*, pp. 535–543, 2015.

[10] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. 30th International Conference on Machine Learning*, pp. 151–159, 2013.

[11] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions–I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[12] S. Wang and W. Chen, "Thompson sampling for combinatorial semi-bandits," in *Proc. 35th International Conference on Machine Learning*, pp. 5114–5122, 2018.

[13] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1746–1778, 2016.

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

[15] A. O. Saritac and C. Tekin, "Combinatorial multi-armed bandit with probabilistically triggered arms: A case with bounded regret," *arXiv preprint arXiv:1707.07443*, 2017.

[16] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson, "Matroid bandits: Fast combinatorial optimization with learning," in *Proc. 30th Conference on Uncertainty in Artificial Intelligence*, pp. 420–429, 2014.

[17] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *Proc. 31st International Conference on Machine Learning*, pp. 100–108, 2014.

[18] Z. Wen, B. Kveton, and A. Ashkan, "Efficient learning in large-scale combinatorial semi-bandits," in *Proc. 32nd International Conference on Machine Learning*, pp. 1113–1122, 2015.

[19] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, 2003.

[20] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan, "Cascading bandits: Learning to rank in the cascade model," in *Proc. 32nd International Conference on Machine Learning*, pp. 767–776, 2015.

[21] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis.* Cambridge University Press, 2005.