Contextual Combinatorial Cascading Bandits

Shuai Li SHUAILI@CSE.CUHK.EDU.HK
The Chinese University of Hong Kong, Hong Kong

Baoxiang Wang BXWANG@CSE.CUHK.EDU.HK
The Chinese University of Hong Kong, Hong Kong

Shengyu Zhang SYZHANG@CSE.CUHK.EDU.HK
The Chinese University of Hong Kong, Hong Kong

Wei Chen WEIC@MICROSOFT.COM
Microsoft Research, Beijing, China

Abstract

We propose the *contextual combinatorial cascading bandits*, a combinatorial online learning game, where at each time step a learning agent is given a set of contextual information, then selects a list of items, and observes stochastic outcomes of a prefix in the selected items by some stopping criterion. In online recommendation, the stopping criterion might be the first item a user selects; in network routing, the stopping criterion might be the first edge blocked in a path. We consider position discounts in the list order, so that the agent's reward is discounted depending on the position where the stopping criterion is met. We design a UCB-type algorithm, C^3-UCB, for this problem, prove an n-step regret bound $\tilde{O}(\sqrt{n})$ in the general setting, and give finer analysis for two special cases. Our work generalizes existing studies in several directions, including contextual information, position discounts, and a more general cascading bandit model. Experiments on synthetic and real datasets demonstrate the advantage of involving contextual information and position discounts.

1. Introduction

Multi-armed bandit (MAB) has been extensively studied in statistics and machine learning. The problem is usually formulated as a system of K base arms whose rewards are random samples from unknown distributions with unknown means. The learning agent pulls one arm every time and tries to minimize the cumulative regret, which is the difference in cumulative rewards between always pulling the best arm in expectation and playing according to the agent's strategy. The problem of MAB has to deal with the trade-off between exploitation (pulling the best empirical arm) and exploration (trying other arms which are not sufficiently pulled).

Recently, stochastic combinatorial bandit started to draw much attention (Gai et al., 2012; Chen et al., 2013; Lin et al., 2014; Gopalan et al., 2014; Chen et al., 2014; Kveton et al., 2014; 2015b;a;c; Lin et al., 2015; Combes et al., 2015b). At every time step, a learning agent chooses a subset of ground items (super arm) under certain combinatorial constraints. There are several different kinds of feedback: (1) bandit feedback, where the learning agent can only obtain the reward of the chosen super arm; (2) semi-bandit feedback, where the learning agent can also obtain the stochastic outcomes of the all base arms constituting the chosen super arm; (3) cascading feedback, where the learning agent can obtain the reward of the chosen super arm and the weights of some base arms in the chosen super arm, according to some problem-specific stopping criterion for observation.

Cascading feedback model fits into many real application scenarios. For example, in online or mobile recommendation, it is a typical practice that an ordered list (instead of a single item) is recommended to a user, who usually goes over the list based on the recommended order and selects one of her interest to click through. Thus it is reasonable to assume that items before the clicked item is of no interest to the user while user's interest to items after the clicked item is unclear. Another example is in networking routing,

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

where the agent needs to select a routing path that is least likely of being blocked. To determine whether a path is blocked, the agent checks from the source node until meeting a blocked edge. The edges before the blocked one are unblocked and the edges after the blocked one are unobserved. In both of the above applications, we observe the feedbacks of a prefix of items in the chosen ordered list. The bandit problem with such cascading feedback has been studied in recent papers (Kveton et al., 2015a;c). In this paper, we generalize their work in several directions to cover more realistic scenarios.

First, we incorporate contextual information into cascading bandits. In online recommendation, contextual information includes various user and item information, and user behaviors in different contexts are different. Therefore utilizing contextual information is crucial for personalized recommendation. In network routing, temporal and spatial contextual information may also be useful to determine better routing paths. Therefore, we incorporate contextual information into the cascading bandit formulation. Second, cascading bandits studied in (Kveton et al., 2015a;c) treats all positions in the cascading list equally, but in applications different positions may bring different rewards. For example, in online recommendation, we usually prefer users to find their interested item in the list as early as possible to increase user satisfaction. In network routing, we may prefer to hit a blocked edge as late as possible. To model these preferences, we introduce position discounts to the cascading bandit, and we show through experiments that incorporating position discounts may significantly improve the learning result. Finally, we generalize the reward functions of (Kveton et al., 2015a;c), which are based on disjunctive and conjunctive binary random variables, to more general non-linear reward functions satisfying monotonicity and Lipschitz continuity conditions. This generalization allows us to cover new realistic scenarios not covered in (Kveton et al., 2015a;c), as exhibited in Section 4.3.

The organization of this paper is as follows: Section 2 compares our work to previous work; Section 3 presents the formulation of *contextual combinatorial cascading bandits* with general reward functions and position discounts; Section 4 gives our algorithm and main results both on general reward functions and two special reward functions; Section 5 demonstrates the experimental results; and Section 6 gives a final conclusion.

2. Related Work

We first discuss several studies most relevant to our work. Table 1 summarizes the different settings of these work, which we will explain in details next. A comparison of regret bounds by different methods is deferred to Section 4.3, after we provide full results of our regret bounds.

	context	cascading	position discount	general reward
CUCB	no	yes	no	yes
C^2UCB	yes	no	no	yes
Comb-Cascade	no	yes	no	no
C^3-UCB (ours)	yes	yes	yes	yes

Table 1. Comparisons of our setting with previous ones

Kveton et al. (2015a) introduced the cascading model to the multi-armed bandit framework with disjunctive objective, where the set of feasible actions form a uniform matroid, and the reward of an action is 1 if there is at least one "good" item in the list. The combinatorial cascading bandits of (Kveton et al., 2015c) (referred to as 'CombCascade' in Table 1) generalized the framework, allowing each feasible action to be a subset of ground items under combinatorial constraints. It studied a problem of combinatorial bandits with conjunctive objective, where each feasible action is a chain of items, and reward is 1 if all items are "good". Our work generalizes these models and involves both contextual information and position discounts.

The experiments of (Kveton et al., 2015a) found that recommending items in increasing order of their UCBs (meaning the items with lower preference come early) has a better performance than in decreasing order of their UCBs. This unnatural result is perhaps because the order of items does not affect the reward and thus putting low preference items first and high preference items in the back may result in more feedbacks in the cascade. The position discounts introduced in this paper makes the recommended set in the decreasing order of UCBs, which is more realistic. Combes et al. (2015a) considered a similar cascading model to that of (Kveton et al., 2015a) with a particular case of contextual information, where a user is recommended with K items selected from a user-related group. Under this particular setting, they considered position discounts. In our setting, users and items are represented by general feature vectors.

Our work considers contextual combinatorial bandits with cascading feedback and nonlinear reward. The paper (Li et al., 2010) studied multi-armed bandit with contextual information, which recommends one item a time. A recent work (Qin et al., 2014) (referred to as 'C^2UCB' in Table 1) introduced contextual combinatorial bandit with semi-bandit feedback and nonlinear reward. Their feedback is more informative than ours. We relax the feedback from semi-bandit to cascading, yet our algorithm achieves the same order of regret bound.

There have been several pieces of work on partial monitor-

ing MAB, but the results are not applicable to our setting. The papers of (Agrawal et al., 1989; Bartók et al., 2012) studied partial monitoring problems but their algorithms are inefficient in our setting. Lin et al. (2014) studied combinatorial partial monitoring problem and their feedback is a fixed (with respect to the chosen action) combination of weights. In our setting, if the chosen action is fixed, the number of observed items varies with the user's behavior, thus the combination constituting feedback is not fixed.

Chen et al. (2016) considered general reward functions and probabilistic triggering of base arms in the stochastic combinatorial semi-bandit model (referred to as 'CUCB' in Table 1). Cascading bandits can be viewed as one type of probabilistic triggering, where the head of the list is deterministically triggered and all the remaining items are probabilistically triggered. However, their work does not consider contextual information, and it is also difficult to incorporate position discounts in a general triggering model.

3. Problem Formulation

We formulate the problem of *contextual combinatorial cascading bandits* as follows. Suppose we have a finite set $E = \{1, ..., L\}$ of L *ground items*, also referred to as *base arms*. Let $\Pi^k = \{(a_1, ..., a_k) : a_1, ..., a_k \in E, a_i \neq a_j \text{ for any } i \neq j\}$ be the set of all k-tuples of distinct items from E; we call each of such tuples an *action* of length k. We will use $|A|$ to denote the length of an action A. Let $\Pi^{\leq K} = \cup_{k=1}^{K} \Pi^k$ denote the set of all actions with length at most K, and let $\mathcal{S} \subseteq \Pi^{\leq K}$ be the set of *feasible actions* with length at most K. As a convention, we always use boldface symbols to represent random variables, and denote $[m] = \{1, \ldots, m\}$.

At time t, feature vectors $x_{t,a} \in \mathbb{R}^{d\times 1}$ with $\|x_{t,a}\|_2 \leq 1$ for every base arm $a \in E$ are revealed to the learning agent. Each feature vector combines the information of the user and the corresponding base arm. For example, suppose the user at time t is characterized by a feature vector u_t and the base arm a has a feature vector v_a, then we can use $x_{t,a} = u_t v_a^\top$, the outer-product of u_t and v_a, as the combined feature vector of base arm a at time t. (See Section 5.2 for an application.) Denote $x_t = (x_{t,a})_{a\in E}$ as all contextual information at time t. Then the learning agent recommends a feasible action $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, ..., \boldsymbol{a}_{|\boldsymbol{A}_t|}^t) \in \mathcal{S}$ to the user. In the cascading feedback models, the user checks from the first item of recommended action and stops at the $\boldsymbol{O}_t$-th item under some stopping criterion. For example, in recommender systems, the stopping criterion might be that the user finds the first attractive item. Then the learning agent observes the weights of first $\boldsymbol{O}_t$ base arms in $\boldsymbol{A}_t$, denoted by $\boldsymbol{w}_t(\boldsymbol{a}_k^t), k \leq \boldsymbol{O}_t$.

Let $\mathcal{H}_t$ denote the history before the learning agent chooses action at time t. Thus $\mathcal{H}_t$ contains feedback information at all time $s < t$, as well as contextual information at time t. Each arm a has a weight $\boldsymbol{w}_t(a)$, representing the "quality" to the user at time t. Given history $\mathcal{H}_t$, we assume that $\boldsymbol{w}_t(a)$'s are mutually independent random variables with expectation

$$\mathbb{E}[\boldsymbol{w}_t(a)|\mathcal{H}_t] = \theta_*^\top x_{t,a}, \tag{1}$$

where θ_* is an unknown d-dimensional vector with the assumption that $\|\theta_*\|_2 \leq 1$ and $0 \leq \theta_*^\top x_{t,a} \leq 1$ for all t, a. Denote by

$$w_{t,a} = \theta_*^\top x_{t,a}$$

the expected weight of base arm a at time t. We assume that each $\boldsymbol{w}_t(a)$ is a random variable with R-sub-Gaussian tail, which means that for all $b \in \mathbb{R}$,

$$\mathbb{E}[\exp(b(\boldsymbol{w}_t(a) - \theta_*^\top x_{t,a}))|\mathcal{H}_t] \leq \exp(b^2R^2/2).$$

Recall that the random variable $\boldsymbol{O}_t$ is the number of observed base arms in $\boldsymbol{A}_t$ and at time t, the agent observes the first $\boldsymbol{O}_t$ items of $\boldsymbol{A}_t$. We say that item a is *observed* if $a = \boldsymbol{a}_k^t$ for some $k \leq \boldsymbol{O}_t$. Thus, $\mathcal{H}_t$ consists of $\{x_s, \boldsymbol{A}_s = (\boldsymbol{a}_1^s, ..., \boldsymbol{a}_{|\boldsymbol{A}_s|}^s), \boldsymbol{O}_s, \boldsymbol{w}_s(\boldsymbol{a}_k^s)\}_{k\leq \boldsymbol{O}_s, s<t}$ and $\{x_t\}$.

We introduce position discounts $\gamma_k \in [0, 1], k \leq K$. For example in website recommendations by a search engine, if the user selects the first website, the learning agent will receive reward 1; and if the user selects the k-th website, the learning agent will receive a discounted reward γ_k.

The reward function on round t is an application dependent function. We assume the expected reward of action A is a function $f(A, w)$ of *expected weights* $w = (w(a))_{a\in E} \in [0, 1]^E$ (at time t, $w = (w_{t,a})_{a\in E} = (\theta_*^\top x_{t,a})_{a\in E}$) and satisfies the following two assumptions.

Monotonicity The expected reward function $f(A, w)$ is a non-decreasing with respect to w: for any $w, w' \in [0, 1]^E$, if $w(a) \leq w'(a)$, we have $f(A, w) \leq f(A, w')$.

Lipschitz continuity The expected reward function $f(A, w)$ is B-Lipschitz continuous with respect to w together with position discount parameters $\gamma_k, k \leq K$. More specifically, for any $w, w' \in [0, 1]^E$, we have

$$|f(A, w) - f(A, w')| \leq B \sum_{k=1}^{|A|} \gamma_k |w(a_k) - w'(a_k)|,$$

where $A = (a_1, \ldots, a_{|A|})$.

The assumption of Lipschitz continuity gives an estimate of changes in the reward function. To obtain a good estimation of the reward function, the position k with large γ_k needs a good estimation of $w(a_k)$. This means the positions with large γ_k are more important to the reward function.

For example in website recommendations, the learning agent will receive the highest reward if the user selects the first item, so the first position is the most important.

Notice that we do not require the knowledge how $f(A, w)$ is defined. We only assume that the agent has access to an oracle $\mathcal{O}_{\mathcal{S}}(w)$ that outputs recommended action A. More precisely, an oracle $\mathcal{O}_{\mathcal{S}}(w)$ is called an α-approximation oracle for some $\alpha \leq 1$, if on given input w, the oracle returns an action $A = \mathcal{O}_{\mathcal{S}}(w) \in \mathcal{S}$ satisfying $f(A, w) \geq \alpha f(A^*, w)$ where $A^* = \text{argmax}_{A \in \mathcal{S}} f(A, w)$. In applications, the oracle $\mathcal{O}_{\mathcal{S}}(w)$ is often what the best offline algorithm can achieve. For this reason, in the regret defined next, we compare our online algorithm with what the best offline algorithm can produce, i.e. αf_t^*.

The α-regret of action A on time t is

$$R^{\alpha}(t, A) = \alpha f_t^* - f(A, w_t),$$

where $f_t^* = f(A_t^*, w_t)$, $A_t^* = \text{argmax}_{A \in \mathcal{S}} f(A, w_t)$ and $w_t = (\theta_*^\top x_{t,a})_{a \in E}$. Our goal is to minimize the α-regret

$$R^{\alpha}(n) = \mathbb{E}[\sum_{t=1}^{n} R^{\alpha}(t, \boldsymbol{A}_t)].$$

4. Algorithms and Results

4.1. Algorithm

Before presenting the algorithm, we explain the main ideas in the design. By Eq.(1), at time s, we have

$$\mathbb{E}[(\gamma_k \boldsymbol{w}_{s,\boldsymbol{a}_k^s}) | \mathcal{H}_s] = \theta_*^\top (\gamma_k x_{s,\boldsymbol{a}_k^s}).$$

To estimate the expected reward, we first estimate θ_*, which can be viewed as a ridge regression problem on samples x and labels $\boldsymbol{w}$. In details, using the ridge regression on data

$$\{(\gamma_k x_{s,\boldsymbol{a}_k^s}, \gamma_k \boldsymbol{w}_{s,\boldsymbol{a}_k^s})\}_{k \in [\boldsymbol{O}_s], s \in [t]},$$

we get an l^2-regularized least-squares estimate of θ_* with regularization parameter $\lambda > 0$:

$$\hat{\theta}_t = (\boldsymbol{X}_t^\top \boldsymbol{X}_t + \lambda I)^{-1} \boldsymbol{X}_t^\top \boldsymbol{Y}_t, \tag{2}$$

where $\boldsymbol{X}_t \in \mathbb{R}^{(\sum_{s=1}^{t} \boldsymbol{O}_s) \times d}$ is the matrix whose rows are $\gamma_k x_{s,\boldsymbol{a}_k^s}^\top$ and $\boldsymbol{Y}_t$ is a column vector whose elements are $\gamma_k \boldsymbol{w}_s(\boldsymbol{a}_k^s)$, $k \in [\boldsymbol{O}_s], s \in [t]$. Let

$$\boldsymbol{V}_t = \boldsymbol{X}_t^\top \boldsymbol{X}_t + \lambda I = \lambda I + \sum_{s=1}^{t} \sum_{k=1}^{\boldsymbol{O}_s} \gamma_k^2 x_{s,\boldsymbol{a}_k^s} x_{s,\boldsymbol{a}_k^s}^\top.$$

Then $\boldsymbol{V}_t \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix. For any symmetric positive definite matrix $V \in \mathbb{R}^{d \times d}$ and any vector $x \in \mathbb{R}^{d \times 1}$, we define the 2-norm of x based on V to be $\|x\|_V = (x^\top V x)^{\frac{1}{2}}$. Next we obtain a good estimate of the difference between $\hat{\theta}_t$ and θ_* by Theorem 2 in (Abbasi-Yadkori et al., 2011), restated as follows.

Lemma 4.1 *(Abbasi-Yadkori et al., 2011) Let*

$$\boldsymbol{\beta}_t(\delta) = R\sqrt{\ln\left(\frac{\det(\boldsymbol{V}_t)}{\lambda^d \delta^2}\right)} + \sqrt{\lambda}. \tag{3}$$

Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t > 0$, we have

$$\|\hat{\theta}_t - \theta_*\|_{\boldsymbol{V}_t} \leq \boldsymbol{\beta}_t(\delta). \tag{4}$$

Thus with high probability, the estimate $\hat{\theta}$ lies in the ellipsoid centered at θ_* with confidence radius $\boldsymbol{\beta}_t(\delta)$ under $\boldsymbol{V}_t$ norm. Building on this, we can define an upper confidence bound of the expected weight for each base arm a by

$$\boldsymbol{U}_t(a) = \min\left\{\hat{\theta}_{t-1}^\top x_{t,a} + \boldsymbol{\beta}_{t-1}(\delta)\|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}}, 1\right\}. \tag{5}$$

The fact that $\boldsymbol{U}_t(a)$ is an upper confidence bound of expected weight $w_{t,a} = \theta_*^\top x_{t,a}$ is proved in the following lemma.

Lemma 4.2 *When Ineq.(4) holds for time $t-1$, we have*

$$0 \leq \boldsymbol{U}_t(a) - w_{t,a} \leq 2\boldsymbol{\beta}_{t-1}(\delta)\|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}}.$$

Proof. Note $w_{t,a} = \theta_*^\top x_{t,a}$. By Hölder's inequality,

$$\begin{aligned}
&\left|\hat{\theta}_{t-1}^\top x_{t,a} - \theta_*^\top x_{t,a}\right| = \left|[\boldsymbol{V}_{t-1}^{1/2}(\hat{\theta}_{t-1} - \theta_*)]^\top (\boldsymbol{V}_{t-1}^{-1/2} x_{t,a})\right| \\
&\leq \|\boldsymbol{V}_{t-1}^{1/2}(\hat{\theta}_{t-1} - \theta_*)\|_2 \|\boldsymbol{V}_{t-1}^{-1/2} x_{t,a}\|_2 \\
&= \|\hat{\theta}_{t-1} - \theta_*\|_{\boldsymbol{V}_{t-1}} \|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}} \\
&\leq \boldsymbol{\beta}_{t-1}(\delta)\|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}}.
\end{aligned}$$

Because $1 - \theta_*^\top x_{t,a} \geq 0$ and

$$\begin{aligned}
0 &\leq (\hat{\theta}_{t-1}^\top x_{t,a} + \boldsymbol{\beta}_{t-1}(\delta)\|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}}) - \theta_*^\top x_{t,a} \\
&\leq 2\boldsymbol{\beta}_{t-1}(\delta)\|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}},
\end{aligned}$$

the claimed result is obtained. □

Based on the above analysis, we design our algorithm as follows. First, the learning agent computes the upper confidence bounds (UCBs) $\boldsymbol{U}_t \in [0, 1]^E$ for the expected weights of all base arms in E. Second, the agent uses the computed UCBs, $\boldsymbol{U}_t$, to select an action $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, ..., \boldsymbol{a}_{|\boldsymbol{A}_t|}^t)$. Third, assume that the user checks from the first base arm in $\boldsymbol{A}_t$ and stops after checking the $\boldsymbol{O}_t$-th base arm and the agents observes $\boldsymbol{w}_t(\boldsymbol{a}_k^t), k \leq \boldsymbol{O}_t$. Then, the learning agent updates $\boldsymbol{V}_t, \boldsymbol{X}_t, \boldsymbol{Y}_t$ in order to get a newer estimate $\hat{\theta}_t$ of θ_* and new confidence radius $\boldsymbol{\beta}_t(\delta)$. Our proposed algorithm, C^3-UCB, is described in **Algorithm 1**. There we use the notation $[A; B]$ to denote the matrix obtained by stacking A and B vertically like $\begin{pmatrix} A \\ B \end{pmatrix}$.

Algorithm 1 C^3-UCB

Parameters:
  $\{\gamma_k \in [0,1]\}_{k\le K}; \delta = \frac{1}{\sqrt{n}}; \lambda \ge C_\gamma = \sum_{k=1}^K \gamma_k^2$
Initialization:
  $\hat{\theta}_0 = 0, \boldsymbol{\beta}_0(\delta) = 1, \boldsymbol{V}_0 = \lambda I, \boldsymbol{X}_0 = \emptyset, \boldsymbol{Y}_0 = \emptyset$
for all $t = 1, 2, \ldots, n$ **do**
  Obtain context $x_{t,a}$ for all $a \in E$
  $\forall a \in E$, compute
    $\boldsymbol{U}_t(a) = \min\{\hat{\theta}_{t-1}^\top x_{t,a} + \boldsymbol{\beta}_{t-1}(\delta)\|x_{t,a}\|_{\boldsymbol{V}_{t-1}^{-1}}, 1\}$
  //Choose action $\boldsymbol{A}_t$ using UCBs $\boldsymbol{U}_t$
  $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, ..., \boldsymbol{a}_{|\boldsymbol{A}_t|}^t) \leftarrow \mathcal{O}_{\mathcal{S}}(\boldsymbol{U}_t)$
  Play $\boldsymbol{A}_t$ and observe $\boldsymbol{O}_t, \boldsymbol{w}_t(\boldsymbol{a}_k^t), k \in [\boldsymbol{O}_t]$
  //Update statistics
  $\boldsymbol{V}_t \leftarrow \boldsymbol{V}_{t-1} + \sum_{k=1}^{\boldsymbol{O}_t} \gamma_k^2 x_{t,\boldsymbol{a}_k^t} x_{t,\boldsymbol{a}_k^t}^\top$
  $\boldsymbol{X}_t \leftarrow [\boldsymbol{X}_{t-1};\ \gamma_1 x_{t,\boldsymbol{a}_1^t}^\top;\ \ldots;\ \gamma_{\boldsymbol{O}_t} x_{t,\boldsymbol{a}_{\boldsymbol{O}_t}^t}^\top]$
  $\boldsymbol{Y}_t \leftarrow [\boldsymbol{Y}_{t-1};\ \gamma_1 \boldsymbol{w}_t(\boldsymbol{a}_1^t);\ \ldots;\ \gamma_{\boldsymbol{O}_t} \boldsymbol{w}_t(\boldsymbol{a}_{\boldsymbol{O}_t}^t)]$
  $\hat{\theta}_t \leftarrow (\boldsymbol{X}_t^\top \boldsymbol{X}_t + \lambda I)^{-1} \boldsymbol{X}_t^\top \boldsymbol{Y}_t$
  $\boldsymbol{\beta}_t(\delta) \leftarrow R\sqrt{\ln(\det(\boldsymbol{V}_t)/(\lambda^d \delta^2))} + \sqrt{\lambda}$
end for t

4.2. Results

4.2.1. General Results

To state our main theorem, we need some definitions. Let $p_{t,A}$ be the *probability of full observation of* A, that is, the probability of observing all base arms of $A = (a_1, \ldots, a_{|A|})$ at time t. Let $p^* = \min_{1\le t\le T} \min_{A\in\mathcal{S}} p_{t,A}$ be the minimal probability that an action could have all base arms observed over all time. The following is the main theorem on the regret achieved by our C^3-UCB algorithm.

Theorem 4.3 *Suppose the expected reward function $f(A, w)$ is a function of expected weights and satisfies the requirements of monotonicity and B-Lipschitz continuity. Then the α-regret of our algorithm, C^3-UCB, satisfies*

$$R^\alpha(n) \le \frac{2\sqrt{2}B}{p^*}\sqrt{nKd\ln(1 + C_\gamma n/(\lambda d))} \cdot \left(R\sqrt{\ln\left[(1 + C_\gamma n/(\lambda d))^d n\right]} + \sqrt{\lambda}\right) + \alpha\sqrt{n} = O\left(\frac{dBR}{p^*}\sqrt{nK}\ln(C_\gamma n)\right), \quad (6)$$

where R is the sub-Gaussian constant and $C_\gamma = \sum_{k=1}^K \gamma_k^2 \le K$.

The regret bound deteriorates as p^* gets smaller. But unfortunately, this reciprocal dependence on p^* is not known to be improvable, even in the special case of disjunctive objective. Indeed, in that case, the bound in Section 2.3 of (Kveton et al., 2015c) also has a reciprocal dependence on p^*. However, in the special case of conjunctive objective, we will give an improved bound.

Proof. (sketch) Assume $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, \ldots, \boldsymbol{a}_{|\boldsymbol{A}_t|}^t)$. We first employ the monotonicity and B-Lipschitz continuity of f to upper bound regret as follows.

$$R^\alpha(t, \boldsymbol{A}_t) \le 2B \sum_{k=1}^{|\boldsymbol{A}_t|} \boldsymbol{\beta}_{t-1}(\delta)\gamma_k \|x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}. \quad (7)$$

As $\boldsymbol{V}_t$ contains only information of observed base arms, our UCB-type algorithm guarantees the sum of the first $\boldsymbol{O}_t$ items in the right of Ineq. (7) to be small, with the leftover sum (from $\boldsymbol{O}_t + 1$ to $|\boldsymbol{A}_t|$) out of control. We cope with this by a reduction to the case that $\boldsymbol{O}_t = |\boldsymbol{A}_t|$.

$$\begin{aligned}
&\mathbb{E}[R^\alpha(t, \boldsymbol{A}_t)|\mathcal{H}_t] \\
=&\mathbb{E}\left[R^\alpha(t, \boldsymbol{A}_t)\mathbb{E}\left[\frac{1}{p_{t,\boldsymbol{A}_t}}\mathbb{1}\{\boldsymbol{O}_t = |\boldsymbol{A}_t|\}\middle|\boldsymbol{A}_t\right]\middle|\mathcal{H}_t\right] \\
\le&\frac{1}{p^*}\mathbb{E}\left[R^\alpha(t, \boldsymbol{A}_t)\mathbb{1}\{\boldsymbol{O}_t = |\boldsymbol{A}_t|\}|\mathcal{H}_t\right]. \quad (8)
\end{aligned}$$

Combining (7) and (8) gives an upper bound of $R^\alpha(n)$ in terms of $\sum_{t=1}^n \sum_{k=1}^{|\boldsymbol{O}_t|} \gamma_k \|x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}$. The next lemma upper bounds this sum of norms, *squared*.

Lemma 4.4 *If $\lambda \ge C_\gamma$, then for any time t,*

$$\sum_{s=1}^t \sum_{k=1}^{\boldsymbol{O}_s} \|\gamma_k x_{s,\boldsymbol{a}_k^s}\|_{\boldsymbol{V}_{s-1}^{-1}}^2 \le 2d\ln\left(1 + C_\gamma t/(\lambda d)\right).$$

The claimed result follows from Cauchy-Schwartz inequality. See Appendix for a complete proof. □

4.2.2. Disjunctive Objective

In the problem of cascading recommendation, when recommended with an ordered list of items $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, ..., \boldsymbol{a}_{|\boldsymbol{A}_t|}^t)$, the user checks the list in that order. The checking process stops if the user selects one item or has checked all items without selecting anyone. The weight of each base arm a at time t, $\boldsymbol{w}_t(a)$, is a $\{0, 1\}$ value indicating whether the user has selected item a or not. Then the random variable $\boldsymbol{O}_t$ satisfies

$$\boldsymbol{O}_t = \begin{cases} k, & \text{if } \boldsymbol{w}_t(\boldsymbol{a}_j^t) = 0, \forall\, j < k \text{ and } \boldsymbol{w}_t(\boldsymbol{a}_k^t) = 1, \\ |\boldsymbol{A}_t|, & \text{if } \boldsymbol{w}_t(\boldsymbol{a}_j^t) = 0,\ \forall\, j \le |\boldsymbol{A}_t|. \end{cases}$$

If the user selects the k-th item, then the learning agent receives a reward $\gamma_k \in [0, 1]$. Usually $\{\gamma_k\}$ are decreasing with k:

$$1 = \gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_K \ge 0.$$

If the user does not select any item, the agent gets no reward. At the end of time step t, the learning agent observes $\boldsymbol{O}_t, \boldsymbol{w}_t(\boldsymbol{a}_k^t), k \le \boldsymbol{O}_t$ and receives a reward $\boldsymbol{r}_t = \boldsymbol{w}_t(\boldsymbol{a}_{\boldsymbol{O}_t}^t)\gamma_{\boldsymbol{O}_t}$. It is not hard to see that $\boldsymbol{r}_t = \bigvee_{k=1}^{|\boldsymbol{A}_t|} \gamma_k \boldsymbol{w}_t(\boldsymbol{a}_k^t)$ is the disjunctive of discounted weights, where we use the notation that $\bigvee_{k=1}^n a_k = \max_{1\le k\le n} a_k$. Notice that the order of $\boldsymbol{A}_t$ affects both the feedback and the reward.

Now let us define a function $f : \mathcal{S} \times [0,1]^E \to [0,1]$ on $A = (a_1, ..., a_{|A|}) \in \mathcal{S}, w = (w(1), ..., w(L))$ by

$$f(A, w) = \sum_{k=1}^{|A|} \gamma_k \prod_{i=1}^{k-1} (1 - w(a_i))w(a_k). \tag{9}$$

It is easily verified that $\boldsymbol{r}_t = f(\boldsymbol{A}_t, \boldsymbol{w}_t)$ and that $\mathbb{E}[\boldsymbol{r}_t] = f(\boldsymbol{A}_t, \theta_*^\top x_t) = f(\boldsymbol{A}_t, w_t)$. So f is the expected reward function and also a function of expected weights. It can be proved that f satisfies the properties of monotonicity and 1-Lipschitz continuity. (All verifications are left in Appendix.) We thus have the following corollary.

Corollary 4.5 *In the problem of cascading recommendation, the expected reward function is defined in Eq. (9), where $1 = \gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_K \ge 0$. Then the α-regret of our algorithm, C^3-UCB, satisfies*

$$\begin{aligned} R^\alpha(n) \le & \frac{2\sqrt{2}}{p^*}\sqrt{nKd\ln(1 + C_\gamma n/(\lambda d))} \\ & \cdot \left(\sqrt{d\ln[(1 + C_\gamma n/(\lambda d))^d n]} + \sqrt{\lambda}\right) + \alpha\sqrt{n} \\ = & O\left(\frac{d}{p^*}\sqrt{nK}\ln(C_\gamma n)\right). \end{aligned} \tag{10}$$

4.2.3. Conjunctive Objective

The above formulation is on the disjunctive objective, for which the user stops once she finds a "good" item. Similarly, we could also consider the case of conjunctive objective, for which the user stops once she finds a "bad" item. In this case, we derive a better result.

The Bernoulli random variable $\boldsymbol{w}_t(a) \in \{0, 1\}$ indicates the weight of item a at time t satisfying Eq. (1). The learning agent observes the first position k in the given action $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, \ldots, \boldsymbol{a}_{|\boldsymbol{A}_t|}^t)$ with $\boldsymbol{w}_t(\boldsymbol{a}_k^t) = 0$. We also consider partial rewards: if the k-th item is the first item with weight 0, then the learning agent receives reward $1 - \gamma_k$; if all items have weight 1, the agent receives reward 1.

The more items reveals, the more reward the agent should receive. This can be used to model scenarios such as online surveys, where questions are adaptively given, and the more questions are observed, the more helpful it is to the survey conductor. Based on this, we assume that

$$1 = \gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_K \ge 0.$$

At the end of time step t, the learning agent observes $\boldsymbol{O}_t, \boldsymbol{w}_t(\boldsymbol{a}_k^t), \forall k \le \boldsymbol{O}_t$ and receives a reward $\boldsymbol{r}_t$:

$$\boldsymbol{r}_t = \begin{cases} 1 - \gamma_k & \text{if } \boldsymbol{w}_t(\boldsymbol{a}_i^t) = 1, \forall i < k, \text{and } \boldsymbol{w}_t(\boldsymbol{a}_k^t) = 0, \\ 1 & \text{if } \boldsymbol{w}_t(\boldsymbol{a}_i^t) = 1, \forall i \le |\boldsymbol{A}_t|. \end{cases}$$

If we define a function $f : \mathcal{S} \times [0,1]^E \to [0,1]$ on $A = (a_1, \ldots, a_{|A|}) \in \mathcal{S}$ and $w = (w(1), \ldots, w(L))$ by

$$f(A, w) = \sum_{k=1}^{|A|}(1 - \gamma_k) \prod_{i=1}^{k-1} w(a_i)(1 - w(a_k)) + \prod_{i=1}^{|A|} w(a_i), \tag{11}$$

then it is not hard to verify that $\boldsymbol{r}_t = f(\boldsymbol{A}_t, \boldsymbol{w}_t)$ and $\mathbb{E}[\boldsymbol{r}_t] = f(\boldsymbol{A}_t, w_t) = f(\boldsymbol{A}_t, \theta_*^\top x_t)$. Next is our result for the conjunctive objective, where the bound in Theorem 4.3 is improved with p^* replaced by αf^*, where $f^* = \min_t f_t^*$ with $f_t^* = \max_{A\in\mathcal{S}} f(A, w_t)$. In network routing problem, the p^* is related with the bad paths, which can make the probability of observing the whole path very small, and f^* is related with good paths, which makes f^* usually large.

Theorem 4.6 *Suppose $1 = \gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_K \ge 1 - \frac{\alpha}{4} f^*$. Then the α-regret of our algorithm, C^3-UCB, for the conjunctive objective satisfies*

$$\begin{aligned} R^\alpha(n) \le & \frac{\sqrt{128}}{\alpha f^*}\sqrt{nKd\ln(1 + C_\gamma n/(\lambda d))} \\ & \cdot \left(\sqrt{d\ln[(1 + C_\gamma n/(\lambda d))^d n]} + \sqrt{\lambda}\right) + \alpha\sqrt{n} \\ = & O\left(\frac{d}{\alpha f^*}\sqrt{nK}\ln(C_\gamma n)\right). \end{aligned} \tag{12}$$

Proof. (Sketch) First, we prove the expected reward function satisfies the properties of monotonicity and Lipschitz continuity. Then we use the following lemma to replace $\boldsymbol{A}_t$ with a prefix of $\boldsymbol{A}_t$.

Lemma 4.7 *Suppose $1 = \gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_K \ge 0$. Let $A = (a_1, ..., a_{|A|})$. For the time t and the conjunctive objective, there exists a prefix B of A such that*

$$p_{t,B} \ge \frac{\alpha}{2} f_t^* - 1 + \gamma_{|B|}, \qquad R^\alpha(t, B) \ge \frac{1}{2} R^\alpha(t, A).$$

Therefore we can get

$$\begin{aligned} & \mathbb{E}[R^\alpha(t, \boldsymbol{A}_t)|\mathcal{H}_t] \\ & \le \frac{8}{\alpha f^*}\mathbb{E}\left[\boldsymbol{\beta}_{t-1}(\delta)\sum_{k=1}^{\boldsymbol{O}_t} \|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}} \middle| \mathcal{H}_t\right]. \end{aligned}$$

Then similar to the proof of Theorem 4.3 and by the Lemma 4.4, we can prove this theorem. □

4.3. Discussions

In the above, we formulate the general setting of contextual combinatorial cascading bandit with position discounts. This setting generalizes several existing results.

The combinatorial cascading bandit (Kveton et al., 2015c) has a setting similar to ours in Section 4.2.3, but theirs has no contextual information or position discounts (i.e. all $\gamma_k = 1$). So our setting in Section 4.2.3 with regret bound $O\left(\frac{d}{f^*}\sqrt{nK}\ln(n)\right)$, which has an additional term $\sqrt{\frac{d\ln(n)}{f^*}}$ compared to their result. The loss is because we use the technique of linear bandits, (Abbasi-Yadkori et al., 2011), to handle the contextual information for a confidence ellipsoid of θ_*. The regret bound of linear bandits has an additional term of $O(\sqrt{d\ln(n)})$ than the standard stochastic MAB bound $O(\sqrt{nd\ln(n)})$. A similar comparison can be made for the disjunctive objective in our Section 4.2.2 and in Section 2.3 of (Kveton et al., 2015c).

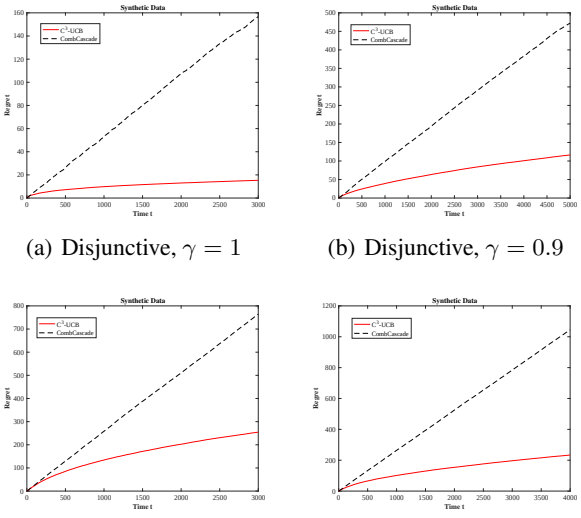
We can also compare to the work (Qin et al., 2014), where there are no triggering and no position discounts, and all base arms in action $\boldsymbol{A}_t$ can be observed. So in their setting, our random variable $\boldsymbol{O}_t$ becomes a deterministic value $|\boldsymbol{A}_t|$ and the probability, $p_{t,\boldsymbol{A}_t}$, is 1. Thus $p^* = 1$. Also all position discounts $\gamma_k = 1$. Then our regret bound of Theorem 4.3 is of the same order with theirs. Notice that the Lipschitz bound C in their setting is $\sqrt{K}$.

Our algorithm also applies to the setting of (Chen et al., 2013) by allowing contextual information. Note the probability p^* in our results should be the probability that all base arms in $\tilde{S}$ (in their notation) will be triggered.

Our setting generalizes several existing works on cascading feedback, and actually covers more cases. Consider in network routing problem, edges have latency following exponential distribution. The observed latency follows cutoff exponential distribution(we will not wait edges to react for arbitrary long time) with mean $\theta_*^\top x$. Suppose we treat an edge as blocked if its latency is larger than some tolerance τ. The edge has reward 1 if it is unblocked; otherwise, the reward is 0. Then the expected reward of an edge is a function of $\theta_*^\top x$, instead of $\theta_*^\top x$ in the conjunctive case. A path has reward 1 only when all its edges have rewards 1. Then it can be proved that the resulting expected reward function, which cannot be represented in conjunctive/disjunctive objectives as in (Kveton et al., 2015a;c), is monotone and satisfies Lipschitz continuity. Details can be found in Appendix.

For the p^* in the result of Theorem 4.3, in our understanding, it is always tied to the reward function. In general, if the probability of observing a base arm i is very small causing $1/p^*$ large while observing i is not tied to the reward function, we can ignore arm i so that it does not affect p^*. In the other extreme, when observing arm i indeed could make a difference in the optimal action selection, one has to observe arm i, no matter how small its observation probability, which means in this case a regret with $1/p^*$ is reasonable. In other cases when p^* is tied with the reward function but not the optimal reward, one may add some condition similar to the one in Lemma 4.7 to detach p^* from the regret.

In addition, the assumption of monotonicity (together with the offline oracle assumption) can be removed if the problem can easily find $A_t = \text{argmax}_{A,w}\{f(A,w)|A,w \text{ in confidence ellipsoid}\}$.



(a) Disjunctive, $\gamma = 1$ (b) Disjunctive, $\gamma = 0.9$

(c) Conjunctive, $\gamma = 1$ (d) Conjunctive, $\gamma = 0.9$

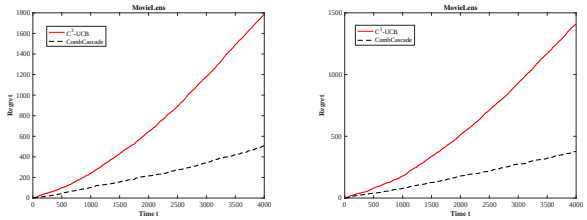
Figure 1. Synthetic Data Set, $L = 200$, $K = 4$, $d = 20$

5. Experiments

We evaluate our algorithm, C^3-UCB, in a synthetic setting and two real applications. The results are compared with *CombCascade*, the study most related to ours, and demonstrate the advantage to involve contextual information and position discounts. In experiments, we set the position discounts γ_k to be γ^{k-1} for some γ.

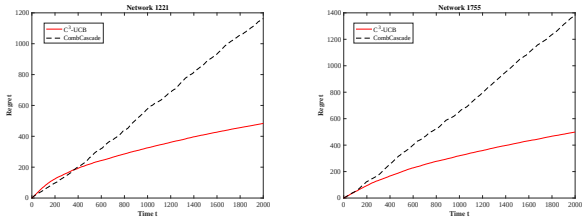
5.1. Synthetic Data

In the first experiment, we compare C^3-UCB to *CombCascade* on synthetic problems. The problem is a contextual cascading bandit with $L = 200$ items and $K = 4$, where at each time t the agent recommends K items to the user. At first, we randomly choose a $\theta \in \mathbb{R}^{d-1}$ with $\|\theta\|_2 = 1$ and let $\theta_* = (\frac{\theta}{2}, \frac{1}{2})$. Then at each time t, we randomly assign $x'_{t,a} \in \mathbb{R}^{d-1}$ with $\|x'_{t,a}\|_2 = 1$ to arm a and use $x_{t,a} = (x'_{t,a}, 1)$ to be the contextual information for



(a) $\gamma = 1$ (b) $\gamma = 0.9$

Figure 2. Rewards on MovieLens, $L = 400$, $K = 4$, $d = 400$



(a) Network 1221, $\gamma = 1$ (b) Network 1755, $\gamma = 0.9$

Figure 3. Network, $d = 5$

arm a. This processing will guarantee the inner product $\theta_*^\top x_{t,a} = \frac{1}{2}(\theta^\top x'_{t,a} + 1) \in [0, 1]$. Next we generate the weight for arm a at time t by a random sample from the Bernoulli distribution with mean $\theta_*^\top x_{t,a}$.

We conduct experiments in four settings. Under each setting, the learning agent chooses a set of K items out of L ground items. The first two are of disjunctive objective where the learning agent observes a prefix of the chosen K items until the first one with weight 1; the last two are of conjunctive objective where the learning agent observes from the first item until the first one with weight 0. Notice that with $\gamma \leq 1$, the oracle selects a set of K items with highest UCBs in their decreasing order. The regrets are shown in Fig. 1 and our algorithm outperforms *CombCascade* algorithm because they do not make use of the contextual information.

5.2. Movie Recommendation

In this experiment, we evaluate C^3-UCB on movie recommendation with data set *MovieLens* (Lam & Herlocker, 2015) of 2015.

The learning problem is formulated as follows. There is a big sparse matrix $A \in \{0, 1\}^{N_1 \times N_2}$ where $A(i, j) = 1$ denotes user i has watched movie j. Next, we split A as $H + F$ by putting 1-entry in A to H or F according to a Bernoulli distribution $\sim \text{Ber}(p)$ for some fixed p. We regard H as known information about history "what users have watched" and regard F as future criterion. We use H to derive feature vectors of both users and movies by SVD decomposition, $H = USM^\top$ where $U = (u_1; ...; u_{N_1})$ and $M = (m_1; ...; m_{N_2})$. At every time t, we randomly choose a user $I_t \in [N_1]$. Then in the same spirit of (Li et al., 2010), we use $x_{t,j} = u_{I_t} m_j^\top$, the outer product of u_{I_t} and m_j, as the contextual information for each movie j. The real weight of movie j at time t, $\boldsymbol{w}_t(j)$, is $F(I_t, m_j)$.

For this experiment, we randomly choose $L = 400$ movies and recommend $K = 4$ movies at each time. We experiment with both $\gamma = 1$ (no position discount) and $\gamma = 0.9$, and compare our algorithm with *CombCascade*. The results are shown in Fig.2. The rewards of our algorithms are 3.52 and 3.736 times of those of *CombCascade* (for $\gamma = 1$ and 0.9, respectively), which demonstrate the advantage to involve contextual information in real applications.

5.3. Network Routing Problem

In this experiment, we evaluate C^3-UCB on network routing problem with *RocketFuel* dataset (Spring et al., 2004).

The ground set E is the set of links in the network. Before learning, the environment randomly chooses a d-dimensional vector $\theta_* \in [0, 1]^d$. At each time t, a pair of source and destination nodes are randomly chosen and the feasible action set $\mathcal{S}_t$ at time t contains all simple paths, paths without cycles, between the source and destination. Any edge a in the set $\mathcal{S}_t$ is assigned with a random d-dimensional contextual information vector $x_{t,a}$. Notice here both θ and x have been processed like in Section 5.1 such that $\theta_*^\top x \in [0, 1]$. The weight for each edge a is a sample from Bernoulli distribution with mean $\theta_*^\top x_{t,a}$. Then the learning agent recommends a feasible path $A = (a_1, ..., a_{|A|})$ between source and destination nodes to maximize the expected reward in the conjunctive objective. We experiment on different position discounts. The regrets are shown in Fig. 3 (a), (b).

6. Conclusions

In this paper, we propose contextual combinatorial cascading bandits with position discounts, where each action is an ordered list and only a prefix of the action is observed each time. We propose a C^3-UCB algorithm and prove a cumulative regret bound for general reward functions and two special reward functions. The experiments conducted demonstrate the advantage to involve contextual information and position discounts. In future, we would like to investigate on lower bounds of the regret and cascading on general graphs.

Acknowledgment

The work was supported in part by Research Grants Council of the Hong Kong S.A.R. (Project No. CUHK419413), Microsoft Research Asia Fund (Project No. FY15-RES-THEME-025), and the National Natural Science Foundation of China (Grant No. 61433014).

References

Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Agrawal, Rajeev, Teneketzis, Demosthenis, and Anantharam, Venkatachalam. Asymptotically efficient adaptive allocation schemes for controlled markov chains: Finite parameter space. *Automatic Control, IEEE Transactions on*, 34(12):1249–1259, 1989.

Bartók, Gábor, Zolghadr, Navid, and Szepesvári, Csaba. An adaptive algorithm for finite stochastic partial monitoring. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Chen, Shouyuan, Lin, Tian, King, Irwin, Lyu, Michael R, and Chen, Wei. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pp. 379–387, 2014.

Chen, Wei, Wang, Yajun, and Yuan, Yang. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 151–159, 2013.

Chen, Wei, Wang, Yajun, Yuan, Yang, and Wang, Qinshi. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.

Combes, Richard, Magureanu, Stefan, Proutiere, Alexandre, and Laroche, Cyrille. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 231–244. ACM, 2015a.

Combes, Richard, Shahi, Mohammad Sadegh Talebi Mazraeh, Proutiere, Alexandre, et al. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pp. 2107–2115, 2015b.

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

Gopalan, Aditya, Mannor, Shie, and Mansour, Yishay. Thompson sampling for complex online problems. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 100–108, 2014.

Kveton, Branislav, Wen, Zheng, Ashkan, Azin, Eydgahi, Hoda, and Eriksson, Brian. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

Kveton, Branislav, Szepesvári, Csaba, Wen, Zheng, and Ashkan, Azin. Cascading bandits: learning to rank in the cascade model. In *Proceedings of the 32th International Conference on Machine Learning*, 2015a.

Kveton, Branislav, Wen, Zheng, Ashkan, Azin, and Szepesvári, Csaba. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics(AISTATS)*, 2015b.

Kveton, Branislav, Wen, Zheng, Ashkan, Azin, and Szepesvari, Csaba. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 2015c.

Lam, Shyong and Herlocker, Jon. Movielens 20m dataset. 2015. URL http://grouplens.org/datasets/movielens/20m/.

Li, Lihong, Chu, Wei, Langford, John, and Schapire, Robert E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. ACM, 2010.

Lin, Tian, Abrahao, Bruno, Kleinberg, Robert, Lui, John, and Chen, Wei. Combinatorial partial monitoring game with linear feedback and its applications. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 901–909, 2014.

Lin, Tian, Li, Jian, and Chen, Wei. Stochastic online greedy learning with semi-bandit feedbacks. In *Advances in Neural Information Processing Systems*, pp. 352–360, 2015.

Qin, Lijing, Chen, Shouyuan, and Zhu, Xiaoyan. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, 2014.

Spring, Neil, Mahajan, Ratul, Wetherall, David, and Anderson, Thomas. Measuring isp topologies with rocketfuel. *Networking, IEEE/ACM Transactions on*, 12(1): 2–16, 2004.

A. Proof of Theorem 4.3

The following lemma provides the important result on the regret at each time t in terms of feature vectors of base arms at the time.

Lemma A.1 *For any time t and $\boldsymbol{A}_t = (\boldsymbol{a}_1^t, \ldots, \boldsymbol{a}_{|A|}^t)$, if f satisfies the two assumptions of monotonicity and B-Lipschitz continuity, then we have*

$$R^\alpha(t, \boldsymbol{A}_t) \leq 2B \sum_{k=1}^{|\boldsymbol{A}_t|} \gamma_k \boldsymbol{\beta}_{t-1}(\delta) \|x_{t,\boldsymbol{a}_k}\|_{\boldsymbol{V}_{t-1}^{-1}}.$$

Proof. By Lemma 4.2, $w_t \leq \boldsymbol{U}_t$. Then

$$f(\boldsymbol{A}_t, \boldsymbol{U}_t) \geq \alpha f(\boldsymbol{A}^{\boldsymbol{U}_t}, \boldsymbol{U}_t) \geq \alpha f(A_t^*, \boldsymbol{U}_t) \geq \alpha f(A_t^*, w_t),$$

where $\boldsymbol{A}^{\boldsymbol{U}_t} = \operatorname{argmax}_{A \in \mathcal{S}} f(A, \boldsymbol{U}_t)$. The first inequality is because we choose $\boldsymbol{A}_t$ by the α-approximation oracle with input $\boldsymbol{U}_t$; the second inequality is because the maximum of $f(\boldsymbol{A}^{\boldsymbol{U}_t}, \boldsymbol{U}_t)$ given $\boldsymbol{U}_t$; the third inequality is by the monotonicity of f and the property $\boldsymbol{U}_t \geq w_t$. Then

$$\begin{aligned} R^\alpha(t, \boldsymbol{A}_t) =& \alpha f(A_t^*, w_t) - f(\boldsymbol{A}_t, w_t) \\ \leq& f(\boldsymbol{A}_t, \boldsymbol{U}_t) - f(\boldsymbol{A}_t, w_t). \end{aligned}$$

By Lipschitz continuity of f and Lemma 4.2,

$$\begin{aligned} f(\boldsymbol{A}_t, \boldsymbol{U}_t) &\leq f(\boldsymbol{A}_t, w_t) + B \sum_{k=1}^{|\boldsymbol{A}_t|} \gamma_k \left| \boldsymbol{U}_t(\boldsymbol{a}_k^t) - w_t(\boldsymbol{a}_k^t) \right| \\ &\leq f(\boldsymbol{A}_t, w_t) + 2B \sum_{k=1}^{|\boldsymbol{A}_t|} \gamma_k \boldsymbol{\beta}_{t-1}(\delta) \|x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}. \end{aligned}$$

Then we have

$$\begin{aligned} R^\alpha(t, A) \leq& f(\boldsymbol{A}_t, \boldsymbol{U}_t) - f(\boldsymbol{A}_t, w_t) \\ \leq& 2B \sum_{k=1}^{|\boldsymbol{A}_t|} \gamma_k \boldsymbol{\beta}_{t-1}(\delta) \|x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}. \end{aligned}$$

□

Notice that the upper bound of $R^\alpha(t, \boldsymbol{A}_t)$ is in terms of all base arms of $\boldsymbol{A}_t$. However, it is hard to estimate an upper bound for $\sum_{t=1}^T \sum_{k=1}^{|\boldsymbol{A}_t|} \|x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}$ because $\boldsymbol{V}_t$ only contains information of observed base arms. Thus we need the following lemma.

Lemma A.2 *Suppose the Ineq.(4) holds for time $t-1$. Then*

$$\mathbb{E}[R^\alpha(t, \boldsymbol{A}_t) | \mathcal{H}_t] \leq \frac{2B}{p^*} \mathbb{E}\left[\boldsymbol{\beta}_{t-1}(\delta) \sum_{k=1}^{\boldsymbol{O}_t} \|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}} \middle| \mathcal{H}_t \right].$$

Proof.

$$\begin{aligned} &\mathbb{E}[R^\alpha(t, \boldsymbol{A}_t) | \mathcal{H}_t] \\ =& \mathbb{E}\left[R^\alpha(t, \boldsymbol{A}_t) \mathbb{E}\left[\frac{1}{p_{t,\boldsymbol{A}_t}} \mathbb{1}\{\boldsymbol{O}_t = |\boldsymbol{A}_t|\} \middle| \boldsymbol{A}_t \right] \middle| \mathcal{H}_t \right] && (13) \\ =& \mathbb{E}\left[R^\alpha(t, \boldsymbol{A}_t) \frac{1}{p_{t,\boldsymbol{A}_t}} \mathbb{1}\{\boldsymbol{O}_t = |\boldsymbol{A}_t|\} \middle| \mathcal{H}_t \right] && (14) \\ \leq& \frac{1}{p^*} \mathbb{E}\left[R^\alpha(t, \boldsymbol{A}_t) \mathbb{1}\{\boldsymbol{O}_t = |\boldsymbol{A}_t|\} \middle| \mathcal{H}_t \right] && (15) \\ \leq& \frac{2B}{p^*} \mathbb{E}\left[\sum_{k=1}^{|\boldsymbol{A}_t|} \gamma_k \boldsymbol{\beta}_{t-1}(\delta) \|x_{t,a_k}\|_{\boldsymbol{V}_{t-1}^{-1}} \mathbb{1}\{\boldsymbol{O}_t = |\boldsymbol{A}_t|\} \middle| \mathcal{H}_t \right] && (16) \\ \leq& \frac{2B}{p^*} \mathbb{E}\left[\sum_{k=1}^{\boldsymbol{O}_t} \gamma_k \boldsymbol{\beta}_{t-1}(\delta) \|x_{t,a_k}\|_{\boldsymbol{V}_{t-1}^{-1}} \middle| \mathcal{H}_t \right]. && (17) \end{aligned}$$

Eq.(13) is because when $\boldsymbol{A}_t$ is fixed, $p_{t,\boldsymbol{A}_t}$ is the probability of $\boldsymbol{O}_t = |\boldsymbol{A}_t|$, and thus the inner expectation is 1. Eq.(14) is because when the history $\mathcal{H}_t$ is fixed, $\boldsymbol{A}_t$ is fixed by the deterministic algorithm C^3-UCB, and thus there is no need to write the conditional expectation. Ineq.(15) is because $p_{t,\boldsymbol{A}_t} \geq p^*$ by definition. Ineq.(16) is by Lemma A.1 and the fact $f(A_t^*, \boldsymbol{U}_t) \leq f(\boldsymbol{A}_t, \boldsymbol{U}_t)$, which is true because algorithm C^3-UCB selects action $\boldsymbol{A}_t$ as the best action with respect to $\boldsymbol{U}_t$. Ineq.(17) is by simply arguing on the two cases for the indicator function $\mathbb{1}\{\boldsymbol{O}_t = |\boldsymbol{A}_t|\}$. □

From the above lemma, we can bound the regret at time t in terms of the observed base arms, for which we further bound in Lemma A.4. But before that, we need the following technical lemma.

Lemma A.3 *Let $x_i \in \mathbb{R}^{d \times 1}, 1 \leq i \leq n$. Then we have*

$$\det\left(I + \sum_{i=1}^n x_i x_i^\top \right) \geq 1 + \sum_{i=1}^n \|x_i\|_2^2.$$

Proof. Denote the eigenvalues of $I + \sum_{i=1}^n x_i x_i^\top$ by $1 + \alpha_1, ..., 1 + \alpha_d$ with $\alpha_j \geq 0, 1 \leq j \leq d$. Then

$$\begin{aligned} &\det(I + \sum_{i=1}^n x_i x_i^\top) = \prod_{j=1}^d (1 + \alpha_j) \\ \geq& 1 + \sum_{j=1}^d \alpha_j = 1 - d + \sum_{i=1}^d (1 + \alpha_i) \\ =& 1 - d + \operatorname{trace}(I + \sum_{i=1}^n x_i x_i^\top) = 1 - d + d + \sum_{i=1}^n \|x_i\|_2^2 \\ =& 1 + \sum_{i=1}^n \|x_i\|_2^2. \end{aligned}$$

□

Lemma A.4 *If $\lambda \geq C_\gamma$, then*

$$\sum_{s=1}^{t}\sum_{k=1}^{\boldsymbol{O}_s} \|\gamma_k x_{s,\boldsymbol{a}_k^s}\|^2_{\boldsymbol{V}_{s-1}^{-1}} \leq 2\ln\left(\frac{\det(\boldsymbol{V}_t)}{\lambda^d}\right).$$

Proof.

$$\begin{aligned}
&\det(\boldsymbol{V}_t) = \det\left(\boldsymbol{V}_{t-1} + \sum_{k=1}^{\boldsymbol{O}_t}(\gamma_k x_{t,\boldsymbol{a}_k^t})(\gamma_k x_{t,\boldsymbol{a}_k^t}^\top)\right)\\
&=\det(\boldsymbol{V}_{t-1})\\
&\quad\cdot\det\left(I + \sum_{k=1}^{\boldsymbol{O}_t}\gamma_k \boldsymbol{V}_{t-1}^{-1/2} x_{t,\boldsymbol{a}_k^t}(\gamma_k \boldsymbol{V}_{t-1}^{-1/2} x_{t,\boldsymbol{a}_k^t})^\top\right) \qquad (18)\\
&\geq \det(\boldsymbol{V}_{t-1})\left(1 + \sum_{k=1}^{\boldsymbol{O}_t}\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|^2_{\boldsymbol{V}_{t-1}^{-1}}\right) \qquad (19)\\
&\geq \det(\lambda I)\prod_{s=1}^{t}\left(1 + \sum_{k=1}^{\boldsymbol{O}_s}\|\gamma_k x_{s,\boldsymbol{a}_k^s}\|^2_{\boldsymbol{V}_{s-1}^{-1}}\right), \qquad (20)
\end{aligned}$$

where Eq.(18) is by the fact that $V + U = V^{1/2}(I + V^{-1/2}UV^{-1/2})V^{1/2}$ for a symmetric positive definite matrix V, Ineq.(19) is due to Lemma A.3, and Ineq.(20) is by repeatedly applying Ineq.(19).

Because

$$\|\gamma_k x_{s,\boldsymbol{a}_k^s}\|^2_{\boldsymbol{V}_{s-1}^{-1}} \leq \|\gamma_k x_{s,\boldsymbol{a}_k^s}\|_2^2/\lambda_{\min}(\boldsymbol{V}_{s-1}) \leq \gamma_k^2/\lambda,$$

where $\lambda_{\min}(\boldsymbol{V}_{s-1})$ is the minimum eigenvalue of $\boldsymbol{V}_{s-1}$, we have

$$\sum_{k=1}^{\boldsymbol{O}_s}\|\gamma_k x_{s,\boldsymbol{a}_k^s}\|^2_{\boldsymbol{V}_{s-1}^{-1}} \leq \frac{1}{\lambda}\sum_{k=1}^{K}\gamma_k^2 = C_\gamma/\lambda \leq 1$$

Using the fact that $2\ln(1+u) \geq u$ for any $u \in [0,1]$, we get

$$\begin{aligned}
&\sum_{s=1}^{t}\sum_{k=1}^{\boldsymbol{O}_s}\|\gamma_k x_{s,\boldsymbol{a}_k^s}\|^2_{\boldsymbol{V}_{s-1}^{-1}}\\
\leq& 2\sum_{s=1}^{t}\ln\left(1 + \sum_{k=1}^{\boldsymbol{O}_s}\|\gamma_k x_{s,\boldsymbol{a}_k^s}\|^2_{\boldsymbol{V}_{s-1}^{-1}}\right)\\
=& 2\ln\prod_{s=1}^{t}\left(1 + \sum_{k=1}^{\boldsymbol{O}_s}\|\gamma_k x_{s,\boldsymbol{a}_k^s}\|^2_{\boldsymbol{V}_{s-1}^{-1}}\right)\\
\leq& 2\ln\left(\frac{\det(\boldsymbol{V}_t)}{\det(\lambda I)}\right),
\end{aligned}$$

where the last inequality is from Ineq.(20). □

Finally the last lemma bounds $\det(\boldsymbol{V}_t)$.

Lemma A.5 *$\det(\boldsymbol{V}_t)$ is increasing with respect to t and*

$$\det(\boldsymbol{V}_t) \leq (\lambda + C_\gamma t/d)^d.$$

Proof. To prove $\det(\boldsymbol{V}_t)$ is increasing with respect to t, it is enough to prove that

$$\det(V + xx^\top) \geq \det(V),$$

for any symmetric positive definite matrix $V \in \mathbb{R}^{d\times d}$ and column vector $x \in \mathbb{R}^{d\times 1}$. In fact,

$$\begin{aligned}
&\det(V + xx^\top)\\
=&\det(V)\det(I + V^{-1/2}xx^\top V^{-1/2})\\
=&\det(V)\det(1 + \|V^{-1/2}x\|^2)\\
\geq&\det(V).
\end{aligned}$$

The second equality above is due to Sylvester's determinant theorem, which states that $\det(I + AB) = \det(I + BA)$.

Let the eigenvalues of $\boldsymbol{V}_t$ be $\lambda_1, \ldots, \lambda_d > 0$. Then

$$\begin{aligned}
\det(\boldsymbol{V}_t) &= \lambda_1 \cdot \cdots \cdot \lambda_d\\
&\leq \left(\frac{\lambda_1 + \ldots + \lambda_d}{d}\right)^d = (\mathrm{trace}(\boldsymbol{V}_t)/d)^d.
\end{aligned}$$

Also,

$$\begin{aligned}
&\mathrm{trace}(\boldsymbol{V}_t)\\
=&\mathrm{trace}(\lambda I) + \sum_{s=1}^{t}\sum_{k=1}^{\boldsymbol{O}_s}\gamma_k^2\mathrm{trace}(x_{s,\boldsymbol{a}_k^s}x_{s,\boldsymbol{a}_k^s}^\top)\\
=&d\lambda + \sum_{s=1}^{t}\sum_{k=1}^{\boldsymbol{O}_s}\gamma_k^2\|x_{s,\boldsymbol{a}_k^s}\|_2^2\\
\leq&d\lambda + \sum_{s=1}^{t}\sum_{k=1}^{K}\gamma_k^2 = d\lambda + tC_\gamma.
\end{aligned}$$

Thus we have $\det(\boldsymbol{V}_t) \leq (\lambda + C_\gamma t/d)^d$. □

Lemma A.4 and Lemma A.5 combined imply Lemma 4.4. Now we are ready to prove the main theorem.

Proof. [of Theorem 4.3] Suppose Ineq.(4) holds for all time t, which is true with probability $1-\delta$. Then with probability $1-\delta$, we have

$$\begin{aligned}
R^\alpha(n) &= \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}[R^\alpha(t,\boldsymbol{A}_t)|\mathcal{H}_t]\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^{n}\frac{2B}{p^*}\mathbb{E}\left[\boldsymbol{\beta}_{t-1}(\delta)\sum_{k=1}^{\boldsymbol{O}_t}\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}\middle|\mathcal{H}_t\right]\right] \qquad (21)
\end{aligned}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{n}\frac{2B}{p^*}\boldsymbol{\beta}_n(\delta)\mathbb{E}\left[\sum_{k=1}^{\boldsymbol{O}_t}\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}\middle|\mathcal{H}_t\right]\right] \quad (22)$$

$$\leq \mathbb{E}\left[\frac{2B}{p^*}\boldsymbol{\beta}_n(\delta)\sqrt{\left(\sum_{t=1}^{n}\boldsymbol{O}_t\right)\left(\sum_{t=1}^{n}\sum_{k=1}^{\boldsymbol{O}_t}\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|^2_{\boldsymbol{V}_{t-1}^{-1}}\right)}\right] \quad (23)$$

$$\leq \mathbb{E}\left[\frac{2B}{p^*}\left(R\sqrt{\ln\left(\frac{\det(\boldsymbol{V}_n)}{\lambda^d/n}\right)}+\sqrt{\lambda}\right)\cdot\sqrt{nK\cdot 2\ln\left(\frac{\det(\boldsymbol{V}_n)}{\lambda^d}\right)}\right] \quad (24)$$

$$\leq \frac{2\sqrt{2}B}{p^*}\left(R\sqrt{\ln[(1+C_\gamma n/(\lambda d))^d n]}+\sqrt{\lambda}\right)\cdot\sqrt{nKd\ln(1+C_\gamma n/(\lambda d))}. \quad (25)$$

Ineq.(21) is by Lemma A.2. Ineq.(22) is because $\boldsymbol{\beta}_t(\delta)$ is increasing with respect to t, derived by the definition of $\boldsymbol{\beta}_t(\delta)$ (Eq.(3)) and Lemma A.5. Ineq.(23) is by the mean inequality. Ineq.(24) is because of the definition of $\boldsymbol{\beta}_t(\delta)$ (Eq.(3)) and Lemma A.4. And the last Ineq.(25) is because estimate of $\det(\boldsymbol{V}_t)$ by Lemma A.5.

Also $R^\alpha(n)\leq\alpha n$ and $\delta=\frac{1}{\sqrt{n}}$. We have

$$R^\alpha(n)\leq\frac{2\sqrt{2}B}{p^*}\left(R\sqrt{\ln[(1+C_\gamma n/(\lambda d))^d n]}+\sqrt{\lambda}\right)\cdot\sqrt{nKd\ln(1+C_\gamma n/(\lambda d))}+\alpha\sqrt{n}.$$

□

B. Proof of Corollary 4.5

In this section, we prove the expected reward function

$$f(A,w)=\sum_{k=1}^{|A|}\gamma_k\prod_{i=1}^{k-1}(1-w(a_i))w(a_k),$$

where $A=(a_1,\ldots,a_{|A|})$, defined in Eq.(9), satisfies the properties of monotonicity and 1-Lipschitz continuity.

Lemma B.1 (monotonicity) *Suppose $1\geq\gamma_1\geq\cdots\geq\gamma_K\geq 0$. Then $f(A,w)$ is increasing with respect to w, that is, if $w\leq w'\in[0,1]^E$, then for any $A\in\prod^{\leq K}$, it holds that*

$$f(A,w)\leq f(A,w').$$

Proof. Without the loss of generality, we prove for the case of $A=(1,\ldots,m)$, where $1\leq m\leq K$. Denote $w=(w_a)_{a\in E}$ and $w'=(w'_a)_{a\in E}$. First, for $1\leq k\leq m$, we have

$$\begin{aligned}&\gamma_k-\sum_{i=k+1}^{m}\gamma_i\prod_{j=k+1}^{i-1}(1-w'_j)w'_i\\ \geq&\gamma_k[1-\sum_{i=k+1}^{m}\prod_{j=k+1}^{i-1}(1-w'_j)w'_i]\\ \geq&\gamma_k\cdot 0=0.\end{aligned}$$

Then it implies that

$$\begin{aligned}&\gamma_k w_k+(1-w_k)\sum_{i=k+1}^{m}\gamma_i\prod_{j=k+1}^{i-1}(1-w'_j)w'_i\\ \leq&\gamma_k w'_k+(1-w'_k)\sum_{i=k+1}^{m}\gamma_i\prod_{j=k+1}^{i-1}(1-w'_j)w'_i.\end{aligned}$$

Therefore,

$$\begin{aligned}&f(A;w_1,\ldots,w_k,w'_{k+1},\ldots,w'_m)\\ =&\sum_{i=1}^{k-1}\gamma_i\prod_{j=1}^{i-1}(1-w_j)w_i+\prod_{j=1}^{k-1}(1-w_j)\\ &\cdot[\gamma_k w_k+(1-w_k)\sum_{i=k+1}^{m}\gamma_i\prod_{j=k+1}^{i-1}(1-w'_j)w'_i]\\ \leq&\sum_{i=1}^{k-1}\gamma_i\prod_{j=1}^{i-1}(1-w_j)w_i+\prod_{j=1}^{k-1}(1-w_j)\\ &\cdot[\gamma_k w'_k+(1-w'_k)\sum_{i=k+1}^{m}\gamma_i\prod_{j=k+1}^{i-1}(1-w'_j)w'_i]\\ =&f(A;w_1,\ldots,w_{k-1},w'_k,\ldots,w'_m).\end{aligned}$$

□

Lemma B.2 (Lipschitz continuity) *Suppose $0\leq\gamma_k\leq 1, k\leq K$. Let $w,w'\in[0,1]^E$. Then*

$$|f(A,w)-f(A,w')|\leq\sum_{k=1}^{|A|}\gamma_k|w(a_k)-w'(a_k)|,$$

where $A=(a_1,\ldots,a_{|A|})$.

Proof. Without the loss of generality, we prove for the case of $A=(1,\ldots,m)$, where $1\leq m\leq K$. Denote $w=(w_a)_{a\in E}$ and $w'=(w'_a)_{a\in E}$. First, suppose $w\leq w'$ and $w'=w+v$.

We prove this by induction. It holds obviously for $m=1$. Suppose it holds for m. Then

$$f(A^{(m+1)},w')$$

$$
\begin{aligned}
=&f(A^{(m)}, w') + \gamma_{m+1} \prod_{k=1}^{m} (1 - w'_k) w'_{m+1} \\
\leq&f(A^{(m)}, w') + \gamma_{m+1} \prod_{k=1}^{m} (1 - w_k)(w_{m+1} + v_{m+1}) \\
\leq&f(A^{(m)}, w) + \sum_{k=1}^{m} \gamma_k v_k \\
&+ \gamma_{m+1} \prod_{k=1}^{m} (1 - w_k) w_{m+1} + \gamma_{m+1} v_{m+1} \\
=&f(A^{(m+1)}, w) + \sum_{k=1}^{m+1} \gamma_k v_k.
\end{aligned}
$$

So it holds for $m+1$.

For general $w, w' \in [0,1]^L$, by Lemma B.1, we have

$$f(A, w \wedge w') \leq \{f(A, w), f(A, w')\} \leq f(A, w \vee w'),$$

where

$$(w \wedge w')_k = \min\{w_k, w'_k\},\ (w \vee w')_k = \max\{w_k, w'_k\}.$$

Then

$$
\begin{aligned}
&|f(A, w) - f(A, w')| \leq f(A, w \vee w') - f(A, w \wedge w') \\
\leq&\sum_{k=1}^{|A|} \gamma_k [(w \vee w')(a_k) - (w \wedge w')(a_k)] \\
=&\sum_{k=1}^{|A|} \gamma_k |w(a_k) - w'(a_k)|
\end{aligned}
$$

□

We have proved the monotonicity and Lipschitz continuity of expected reward function f. Then the corollary follows Theorem 4.3.

C. Proof of Theorem 4.6

In this section, we prove Theorem 4.6. Recall the expected reward function is

$$f(A, w) = \sum_{k=1}^{|A|} (1 - \gamma_k) \prod_{i=1}^{k-1} w(a_i)(1 - w(a_k)) + \prod_{i=1}^{|A|} w(a_i),$$

where $A = (a_1, \ldots, a_{|A|})$, as defined in Eq.(11). First, similar to last section, we prove this reward function satisfies the properties of monotonicity and 1-Lipschitz continuity.

Lemma C.1 (monotonicity) *Suppose $1 \geq \gamma_1 \geq \cdots \geq \gamma_K \geq 0$. Then $f(A, w)$ is increasing with respect to w; that is, if $w \leq w' \in [0,1]^E$, then for any $A \in \prod^{\leq K}$, it holds that*

$$f(A, w) \leq f(A, w').$$

Proof. Denote $A = (a_1, \ldots, a_{|A|})$. Because

$$
\begin{aligned}
&1 - f(A, 1 - w) \\
=&1 - \sum_{k=1}^{|A|} (1 - \gamma_k) \prod_{i=1}^{k-1} (1 - w(a_i)) w(a_k) - \prod_{i=1}^{|A|} (1 - w(a_i)) \\
=&\sum_{k=1}^{|A|} \gamma_k \prod_{i=1}^{k-1} (1 - w(a_i)) w(a_k),
\end{aligned}
$$

then by the proof of Lemma B.1 and $1 - f(A, 1 - w)$ is increasing in w, we can get f is increasing in w. □

Lemma C.2 (Lipschitz continuity) *Suppose $0 \leq \gamma_k \leq 1, k \leq K$. Let $w, w' \in [0,1]^L$. Then*

$$|f(A, w) - f(A, w')| \leq \sum_{k=1}^{|A|} \gamma_k |w(a_k) - w'(a_k)|,$$

where $A = (a_1, \ldots, a_{|A|})$.

Proof. Denote $w = (w_a)_{a \in E}$ and $w' = (w'_a)_{a \in E}$. Similar to the proof of Lemma B.2. It is enough to prove when $w \leq w', w' = w + v$ and $A = (1, \ldots, m), m \leq K$.

We prove this by induction. It holds obviously for $m = 1$. Suppose it holds for m. Then

$$
\begin{aligned}
&f(A^{(m+1)}, w') \\
=&f(A^{(m)}, w') - \prod_{k=1}^{m} w'_k \\
&+ (1 - \gamma_{m+1})(\prod_{k=1}^{m} w'_k)(1 - w'_{m+1}) + \prod_{k=1}^{m+1} w'_k \\
=&f(A^{(m)}, w') - \gamma_{m+1}(\prod_{k=1}^{m} w'_k)(1 - w'_{m+1}) \\
\leq&f(A^{(m)}, w') - \gamma_{m+1}(\prod_{k=1}^{m} w_k)(1 - w'_{m+1}) \\
\leq&f(A^{(m)}, w') - \gamma_{m+1}(\prod_{k=1}^{m} w_k)(1 - w_{m+1} - v_{m+1}) \\
\leq&(f(A^{(m)}, w) + \sum_{k=1}^{m} \gamma_k v_k) \\
&- \gamma_{m+1}(\prod_{k=1}^{m} w_k)(1 - w_{m+1}) + \gamma_{m+1} v_{m+1} \\
\leq&(A^{(m+1)}, w) + \sum_{k=1}^{m+1} \gamma_k v_k.
\end{aligned}
$$

□

The next lemma provides some properties about a prefix of action A in the conjunctive case, which leads to the finding

of a prefix B of A with similar regret and probability of full observation as those of A, as given in Lemma C.4.

Lemma C.3 *Suppose $1 = \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_K \geq 0$. Let $A = (a_1, ..., a_{|A|})$. Let $B_k = (a_1, ..., a_k), k \leq |A|$ be a prefix of A. The expected weights are denoted by w. The probability of full observation of A, p_A, can be formulated as*

$$p_A = \prod_{k=1}^{|A|-1} w_{a_k}.$$

Then for the problem with conjunctive objective and $k < |A|$, we have the following properties:

(1) $f(A, w) \leq (1 - \gamma_{|A|}) + \gamma_{|A|} p_A$;

(2) $f(B_{k+1}, w) \leq f(B_k, w)$;

(3) $p_{B_{k+1}} \leq p_{B_k}$;

(4) $f(B_k, w) \leq (1 - \gamma_{k+1}) + \gamma_{k+1} p_{B_{k+1}}$.

Proof. Let $m = |A|$. By the definition of $f(A, w)$ in Eq. (11), we have

(1)

$$\begin{aligned}
&f(A, w) \\
=&\sum_{k=1}^{|A|} (1 - \gamma_k) \prod_{i=1}^{k-1} w(a_i)(1 - w(a_k)) + \prod_{i=1}^{|A|} w(a_i) \\
\leq&(1 - \gamma_m)(1 - p_A w_m) + p_A w_m \\
\leq&(1 - \gamma_m) + \gamma_m p_A;
\end{aligned}$$

(2)

$$\begin{aligned}
&f(B_k, w) - f(B_{k+1}, w) \\
=&\prod_{i=1}^{k} w_i - (1 - \gamma_{k+1})(\prod_{i=1}^{k} w_i)(1 - w_{k+1}) \\
&- (\prod_{i=1}^{k} w_i) w_{k+1} \\
=&\gamma_{k+1}(\prod_{i=1}^{k} w_i)(1 - w_{k+1}) \geq 0;
\end{aligned}$$

(3) $p_{B_{k+1}} = \prod_{i=1}^{k} w_i \leq \prod_{i=1}^{k-1} w_i = p_{B_k}$;

(4)

$$\begin{aligned}
f(B_k, w) &\leq (1 - \gamma_k)(1 - p_{B_{k+1}}) + p_{B_{k+1}} \\
&\leq (1 - \gamma_{k+1})(1 - p_{B_{k+1}}) + p_{B_{k+1}} \\
&= (1 - \gamma_{k+1}) + \gamma_{k+1} p_{B_{k+1}}
\end{aligned}$$

□

Lemma C.4 *Suppose $1 = \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_K \geq 0$. Let $A = (a_1, ..., a_{|A|})$. For the time t and the conjunctive objective, there exists a prefix B of A such that*

$$p_{t,B} \geq \frac{\alpha}{2} f_t^* - 1 + \gamma_{|B|}, \qquad R^\alpha(t, B) \geq \frac{1}{2} R^\alpha(t, A).$$

Proof. Recall $w_t = \theta_*^\top x_t$. If $f(A, w_t) \geq \frac{\alpha}{2} f_t^*$, then by Lemma C.3 (1),

$$p_{t,A} \geq \frac{\alpha}{2} f_t^* - 1 + \gamma_{|A|}.$$

In this case, we set prefix $B = A$.

Now suppose $f(A, w_t) \leq \frac{\alpha}{2} f_t^*$. Let

$$x_k = f(B_k, w), \;\; y_k = 1 - \gamma_k + \gamma_k p_{t,B_k}, \;\; I_k = [x_k, y_k]$$

Then by Lemma C.3, we have $x_k \leq y_k$, $x_{k+1} \leq x_k \leq y_{k+1}$. Therefore, I_k is indeed an interval and $I_k \cap I_{k+1} \neq \emptyset$. Also, $x_{|A|} = f(A, w)$ and $y_1 = 1$. Thus

$$[f(A, w), 1] = \bigcup_{k=1}^{|A|} I_k.$$

Then there exists a k such that $\frac{\alpha}{2} f_t^* \in I_k$:

$$f(B_k, w) \leq \frac{\alpha}{2} f_t^* \leq 1 - \gamma_k + \gamma_k p_{t,B_k}$$

Then

$$R^\alpha(t, B_k) = \alpha f_t^* - f(B_k, w_t) \geq \frac{\alpha}{2} f_t^* \geq \frac{1}{2} R^\alpha(A, w_t),$$

and

$$p_{t,B_k} \geq \frac{\alpha}{2} f_t^* - 1 + \gamma_{|B_k|}.$$

□

The key difference from the proof of Theorem 4.3 is the following lemma. This lemma corresponds to Lemma A.2 and uses f^* to replace p^*.

Lemma C.5 *If we have $\gamma_K \geq 1 - \frac{\alpha}{4} f^*$, where $f^* = \min_{1 \leq t \leq T} f_t^*$, then*

$$\begin{aligned}
&\mathbb{E}[R^\alpha(t, \boldsymbol{A}_t) | \mathcal{H}_t] \\
&\leq \frac{8}{\alpha f^*} \mathbb{E}\left[\boldsymbol{\beta}_{t-1}(\delta) \sum_{k=1}^{\boldsymbol{O}_t} \| \gamma_k x_{t, \boldsymbol{a}_k^t} \|_{\boldsymbol{V}_{t-1}^{-1}} \middle| \mathcal{H}_t \right].
\end{aligned}$$

Proof. By Lemma C.4 there exists a prefix $\boldsymbol{B}_t$ of $\boldsymbol{A}_t$ such that $p_{t,\boldsymbol{B}_t} \geq \frac{\alpha}{2} f_t^* - 1 + \gamma_{|\boldsymbol{B}_t|}$ and $R(t, \boldsymbol{B}_t) \geq \frac{1}{2} R(t, \boldsymbol{A}_t)$.

By the assumption that $1-\gamma_{|\boldsymbol{B}_t|} \le 1-\gamma_K \le \frac{\alpha}{4} f^* \le \frac{\alpha}{4} f_t^*$, we have $p_{t,\boldsymbol{B}_t} \ge \frac{\alpha}{4} f_t^*$.

Then, similar to the derivation from Eq.(13) to Eq.(15), we have

$$\begin{aligned}
&\mathbb{E}[R^\alpha(t,\boldsymbol{A}_t)|\mathcal{H}_t] \le 2\mathbb{E}[R^\alpha(t,\boldsymbol{B}_t)\,|\,\mathcal{H}_t] \\
=&2\mathbb{E}\left[R^\alpha(t,\boldsymbol{B}_t)\mathbb{E}\left[\frac{1}{p_{t,\boldsymbol{B}_t}}\mathbb{1}\{\boldsymbol{O}_t \ge |\boldsymbol{B}_t|\}\middle|\boldsymbol{B}_t\right]\middle|\mathcal{H}_t\right] \\
=&2\mathbb{E}\left[R^\alpha(t,\boldsymbol{B}_t)\frac{1}{p_{t,\boldsymbol{B}_t}}\mathbb{1}\{\boldsymbol{O}_t \ge |\boldsymbol{B}_t|\}\middle|\mathcal{H}_t\right] \\
\le&\frac{8}{\alpha f_t^*}\mathbb{E}\left[R^\alpha(t,\boldsymbol{B}_t)\mathbb{1}\{\boldsymbol{O}_t \ge |\boldsymbol{B}_t|\}\,|\,\mathcal{H}_t\right]. \qquad (26)
\end{aligned}$$

Next, we have

$$f(A_t^*, w_t) \le f(A_t^*, \boldsymbol{U}_t) \le f(\boldsymbol{A}_t, \boldsymbol{U}_t) \le f(\boldsymbol{B}_t, \boldsymbol{U}_t), \qquad (27)$$

where the first inequality is by Lemmas 4.2 and C.1, the second inequality is by the C^3-UCB algorithm, in which the best action $\boldsymbol{A}_t$ is selected with respect to $\boldsymbol{U}_t$, and the last inequality is by Lemma C.3 (2).

Then by Lemma C.2,

$$f(\boldsymbol{B}_t, \boldsymbol{U}_t) \le f(\boldsymbol{B}_t, w_t) + \sum_{k=1}^{|\boldsymbol{B}_t|} \gamma_k \boldsymbol{\beta}_{t-1}(\delta)\|x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}. \qquad (28)$$

Therefore, combining Eq.(27) and (28), we have

$$\begin{aligned}
R^\alpha(t,\boldsymbol{B}_t) &= f_t^* - f(\boldsymbol{B}_t, \theta_*^\top x_t) \\
&\le \sum_{k=1}^{|\boldsymbol{B}_t|} \boldsymbol{\beta}_{t-1}(\delta)\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}. \qquad (29)
\end{aligned}$$

Finally from Ineq. (26) and (29) we can derive the result

$$\begin{aligned}
&\mathbb{E}_t[R^\alpha(t,\boldsymbol{A}_t)] \\
\le&\frac{8}{\alpha f_t^*}\mathbb{E}\left[\mathbb{1}\{\boldsymbol{O}_t \ge |\boldsymbol{B}_t|\}\sum_{k=1}^{|\boldsymbol{B}_t|}\boldsymbol{\beta}_{t-1}(\delta)\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}\right] \\
\le&\frac{8}{\alpha f^*}\mathbb{E}\left[\sum_{k=1}^{\boldsymbol{O}_t}\boldsymbol{\beta}_{t-1}(\delta)\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}\right].
\end{aligned}$$

□

Then we have the proof of Theorem 4.6.

Proof. [of Theorem 4.6] Similar to the proof of Theorem 4.3, we have the following derivation. Suppose Ineq.(4) holds for all time t, then

$$\begin{aligned}
R^\alpha(n) &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}_t[R^\alpha(t,\boldsymbol{A}_t)]\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^n \frac{8}{\alpha f^*}\mathbb{E}_t\left[\boldsymbol{\beta}_{t-1}(\delta)\sum_{k=1}^{\boldsymbol{O}_t}\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}\right]\right] \qquad (30) \\
&\le \mathbb{E}\left[\frac{8}{\alpha f^*}\boldsymbol{\beta}_n(\delta)\sum_{t=1}^n\sum_{k=1}^{\boldsymbol{O}_t}\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|_{\boldsymbol{V}_{t-1}^{-1}}\right] \qquad (31) \\
&\le \mathbb{E}\left[\frac{8}{\alpha f^*}\boldsymbol{\beta}_n(\delta)\sqrt{\left(\sum_{t=1}^n \boldsymbol{O}_t\right)\mathbb{E}_t\left[\sum_{t=1}^n\sum_{k=1}^{\boldsymbol{O}_t}\|\gamma_k x_{t,\boldsymbol{a}_k^t}\|^2_{\boldsymbol{V}_{t-1}^{-1}}\right]}\right] \qquad (32) \\
&\le \mathbb{E}\left[\frac{\sqrt{128}}{\alpha f^*}\left(\sqrt{\ln\left(\frac{\det(\boldsymbol{V}_n)}{\lambda^d/n}\right)}+\sqrt{\lambda}\right)\right. \\
&\qquad \left.\cdot\sqrt{nK\cdot 2\ln\left(\frac{\det(\boldsymbol{V}_n)}{\lambda^d}\right)}\right] \qquad (33) \\
&\le \frac{\sqrt{128}}{\alpha f^*}\left(\sqrt{\ln[(1+C_\gamma n/(\lambda d))^d n]}+\sqrt{\lambda}\right) \\
&\qquad \cdot\sqrt{nKd\ln(1+C_\gamma n/(\lambda d))}. \qquad (34)
\end{aligned}$$

By Lemma 4.1, $R^\alpha(n) \le \alpha n$ and $\delta = \frac{1}{\sqrt{n}}$, we have

$$\begin{aligned}
R^\alpha(n) \le& \frac{\sqrt{128}}{\alpha f^*}\left(\sqrt{\ln[(1+C_\gamma n/(\lambda d))^d n]}+\sqrt{\lambda}\right) \\
&\cdot\sqrt{nKd\ln(1+C_\gamma n/(\lambda d))}+\alpha\sqrt{n}.
\end{aligned}$$

□

D. The Network Delay Example

In this section, we will give a network with delay example which is an example of general reward function and cannot be covered by the disjunctive objective and conjunctive objective. We also conduct an experiment on this problem.

Suppose there is a network and the latency on each edge of a network is a random variable of an exponential distribution. If the latency is larger than some tolerance τ, then the edge is regarded as *blocked*; if the latency is less than the tolerance τ, then the edge is regarded as *unblocked*. For a path, only when all edges on it are unblocked, the path is unblocked and has reward 1. We want to find the path with largest probability of being unblocked. In addition, since we will only record latency up to some bound b, the observed latency is a random variable of a cut-off exponential distribution. Obviously, we have $\tau \le b$.

Recall the probability density function (PDF) of an exponential random variable X with parameter λ is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

Then the cut-off random variable X_b of X with boundary b is

$$X_b = \begin{cases} X, & 0 \le X \le b \\ b, & X > b. \end{cases}$$

Then

$$\mathbb{P}(X_b = b) = \int_b^\infty \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_b^\infty = e^{-\lambda b}$$

and the mean of the cut-off exponential distribution is

$$\begin{aligned} &\int_0^b x\lambda e^{-\lambda x} dx + b e^{-\lambda b} \\ =&\int_0^b x d(-e^{-\lambda x}) + \tau_1 e^{-\lambda b} \\ =&-xe^{-\lambda x}\Big|_0^b + \int_0^b -e^{-\lambda x} dx + b e^{-\lambda b} \\ =&-b e^{-\lambda b} + (-\frac{1}{\lambda}\, e^{-\lambda x}\Big|_0^b) + \tau_1 e^{-\lambda b} \\ =&\frac{1}{\lambda} - \frac{1}{\lambda} e^{-\lambda b}. \end{aligned}$$

If we denote the mean of X by w_0, then $\lambda = \frac{1}{w_0}$ the mean of cut-off exponential distribution is

$$g(w_0) = w_0(1 - e^{-b/w_0}).$$

And the probability of an edge being blocked is

$$\int_\tau^\infty \frac{1}{w_0} e^{-x/w_0} dx = e^{-\tau/w_0}.$$

So the probability of an edge being unblocked is $1 - e^{-\tau/w_0}$. Therefore, the expected reward of an edge is

$$h(w_0) = 1 - e^{-\tau/w_0}.$$

Denote the mean of the observed latency, X_b, by w. We assume $w \in (0, 1)$. Then the probability of an edge being unblocked is

$$h \circ g^{-1}(w).$$

First,

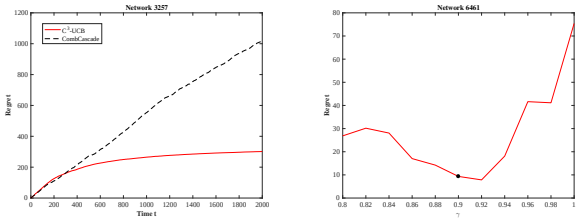
$$g'(w_0) = 1 - e^{-b/w_0} + w(-e^{-b/w_0}\frac{b}{w_0^2}) = 1 - (\frac{b}{w_0} + 1)e^{-b/w_0}.$$

Let $x = \frac{b}{w_0} \in (b, \infty)$, then

$$g'(w_0) = 1 - (x+1)e^{-x} := g_1(x) > 0,$$

And

$$g_1'(x) = -e^{-x} + (x+1)e^{-x} = xe^{-x} > 0.$$



(a) Rewards on network 3257 with latency, $\gamma = 0.9$ (b) Comparison of different γ with true $\gamma^* = 0.9$, $n = 10000$

Figure 4. Network, $d = 5$

So g is an increasing function and satisfies

$$g'(w_0) \in (1 - (b+1)e^{-b}, 1). \tag{35}$$

Second,

$$h'(w_0) = -e^{-\tau/w_0}\frac{\tau}{w_0^2} = -e^{-\tau x}\tau x^2 := h_1(x) < 0,$$

if we denote $x = \frac{1}{w_0} \in (1, \infty)$. Also

$$h_1'(x) = -(-\tau)e^{-\tau x}\tau x^2 - e^{-\tau x}\tau 2x = e^{-\tau x}\tau x(\tau x - 2).$$

If in addition we assume $\tau < 1$, then h is a decreasing function and satisfies

$$-e^{-2}\frac{4}{\tau} \le h'(w_0) \le 0.$$

Therefore, $h \circ g^{-1}$ is monotone decreasing and satisfies Lipschitz continuity with bound

$$B = \frac{4e^b}{e^2(e^b - b - 1)\tau}.$$

Then for each path A, the expected reward of A under weights w, which is the mean of observed latency in this example, is

$$f(A, w) = \prod_{a \in A} (h \circ g^{-1})(w_a). \tag{36}$$

Then it is easy to prove that f is monotone decreasing and satisfies Lipschitz continuity with bound B.

Next is the experiment conducted on the network with latency, where the latency on an edge is a cut-off exponential random variable with mean $\theta_*^\top x$ and the reward is 1 if the latency is less than a tolerance $\tau = 0.8$. The comparison of the cumulative rewards of our algorithm with *CombCascade* is shown in Fig. 4 (a).

In addition, we also conduct an experiment on the influence of γ to the regret. In recommendations, the position

discounts represent users' satisfaction and the learning agent may not know its true value. Suppose the true position discount $\gamma^* = 0.9$, which is used to evaluating reward and regret, and the learning agent sets γ by her experience, which is used to select actions. We experiment on different $\gamma \in \{0.8, 0.82, \ldots, 0.98, 1.0\}$ with the true criterion $\gamma^* = 0.9$. The regrets with different γ's are illustrated in Fig. 4(b), from which it can be seen that the regret is minimized around the true γ^*. The big gap between the regret at $\gamma = 1$ and $\gamma^* = 0.9$ shows that it significantly benefits by exploiting position discounts in the model when such discounts indeed exist in applications.