Collaborative Contextual Combinatorial Cascading Thompson Sampling

Zhenyu Zhu, Liusheng Huang(✉), and Hongli Xu

University of Science and Technology of China, Hefei, Anhui Province, China
zzy7758@mail.ustc.edu.cn
{lshuang,honglixu}@ustc.edu.cn

Abstract. We design and analyze collaborative contextual combinatorial cascading Thompson sampling (C^4-TS). C^4-TS is a Bayesian heuristic to address the cascading bandit problem in the collaborative environment. C^4-TS utilizes posterior sampling strategy to balance the exploration-exploitation tradeoff and it also incorporates the collaborative effect to share information across similar users. Utilizing these two novel features, we prove that the regret upper bound for C^4-TS is $\tilde{O}(d(u + \sqrt{mKT}))$, where d is the dimension of the feature space, u is the number of users, m is the number of clusters, K is the length of the recommended list and T is the time horizon. This regret upper bound matches the theoretical guarantee for UCB-like algorithm in the same settings. We also conduct a set of simulations comparing C^4-TS with the state-of-the-art algorithms. The empirical results demonstrate the advantage of our algorithm over existing works.

1 Introduction

Most recommendation systems recommend an ordered list of candidate items to users due to the limited space. The user examines the recommended list sequentially, clicks on the first satisfying item and stops examining further. The click of the user reveals that the items before the clicked item are not satisfying and the items after the clicked item are unexamined. The recommendation systems observe this feedback and adjust its recommendation strategy accordingly. This kind of interaction is often formulated as the cascading model, which is simple, intuitive and effective in characterizing user behaviours.

We consider the contextual combinatorial cascading model in a collaborative environment. In the stochastic contextual settings, the expected reward of an item is assumed to be a linear function of the item features and a stationary but unknown user vector. At each time step, the learning agent recommends a combination of items to the user. It then observes the cascading feedback of the user and adjusts its recommendation strategy accordingly. The goal of the learning agent is to maximize the cumulative reward in T rounds. As the expected rewards of the items are unknown to the learning agent, it has to balance between exploring new information to improve future performance and exploiting the best empirical items so far. This tradeoff is modelled by the bandit problems which have been well studied in the literature. While effective, standard

X. Wang et al. (Eds.): CollaborateCom 2019, LNICST 292, pp. 118–132, 2019.
https://doi.org/10.1007/978-3-030-30146-0_9

bandit algorithms often work in a content-dependent regime, so that any collaborative effects among users are ignored. This drawback hinders the practical deployment of bandit algorithms in highly dynamic and large-scale domains, in which incorporating collaborative effects often helps to accumulate information more efficiently. Thus, exploiting collaborative effects into bandit algorithms can be one of the most promising approaches to further improvement of the recommendation performance. But it also raises new challenges in the design and analysis of the algorithm.

In this paper, we propose collaborative contextual combinatorial cascading Thompson sampling (C^4-TS) algorithm. Following the approaches in [8,13], C^4-TS maintains a dynamic graph to represent the partition of users. If two users are connected, they are considered to be in the same cluster. The graph is fully-connected at the beginning, and the edges are gradually removed as the algorithm accumulates more information about user preference. At each round, the algorithm considers both the historical feedbacks of the user and the collaborative information to make decisions. It applies posterior matching strategy by recommending the items according to their probability of being optimal. The feedback of the users is then used to update the user vector and the graph.

Our algorithm is based on Thompson sampling because of its advantage over UCB in both empirical performance [5,6,15,17,18] and computational efficiency [3,16]. Although the regret upper bound of UCB-like algorithm in similar settings has been studied [13], the randomness of Thompson sampling presents additional challenges. Under some reasonable assumptions, we utilize the matrix martingale theory to bound the variance of the reward estimator and quantify the exploration-exploitation tradeoff. We prove an upper bound of $\tilde{O}(d(u + \sqrt{mKT}))$ for the expected cumulative regret, where u is the number of users, m is the number of clusters, d is the dimension of feature space, K is the length of the recommended list, and T is the time horizon. The notation $\tilde{O}$ ignores dependence on the logarithmic factors. This bound matches the regret upper bound for UCB-like algorithm. We also conduct experiments on a synthetic dataset to demonstrate the advantage of the model and algorithm over existing studies.

The rest of this paper is organized as follows. Section 2 introduces the related works in similar settings. Section 3 introduces the basic model settings (learning model, notations and assumptions) and presents a detailed description of C^4-TS algorithm. Section 4 provides the theoretical analysis of its regret bound. Section 5 reports the result of simulations. Section 6 concludes this paper.

2 Related Work

Cascading bandit was first introduced by Kveton and Branislav [10]. They also proposed CascadeUCB1 and CascadeKL-UCB to solve the problem and provided gap-dependent regret upper bound of the algorithms. The regret upper bound of CascadeKL-UCB matches the lower bound of the problem within a logarithmic factor. Zong and Ni [18] then generalized the cascading bandit with linear payoff and proposed CascadeLinUCB and CascadeLinTS. They also provided an upper bound on the regret of CascadeLinUCB and suggested that the same theoretical guarantee should hold for CascadeLinTS. The work [12] by Li and Wang generalized the contextual combinatorial cascading setting with position discounts and more general reward functions and

provided a similar theoretical guarantee. The first theoretical analysis of Thompson sampling for non-contextual cascading bandit is provided by Cheung and Tan [6]. They proved that the regret upper bound of CascadeTS matches the state-of-the-art regret bounds for UCB-like algorithms.

Beyond the general settings of cascading bandit and Thompson sampling, our work is also closely related to the dynamic clustering of bandits. Clustering over bandits to utilize collaborative information has been studied in a series of works. These works are based on the assumption that the algorithm servers a large set of users and these users can be partitioned into several groups. All users in the same group can share feedbacks to facilitate customizing personal recommendation. The work [8] first considered online clustering of contextual bandits. It used the confidence interval of the user vector to estimate user similarity and share information across similar users. The work [14] incorporates dynamic clustering to divide users into groups and customizes the bandits to each group. They first used the K-means clustering algorithm within the contextual bandit framework. In [19], the authors developed a collaborative contextual bandit algorithm and leveraged the adjacency graph to share information and feedbacks among similar users while online updating. In [11], the authors extended the work [19] by performing online clustering at both the user side and the item side. They also used a sparse graph to represent the clusters to avoid expensive computation. The work [7] considered a variant of online clustering where the clusters over users are estimated in a context-dependent manner.

The most similar work to ours is [13]. In this paper, the authors first formulated the problem of dynamic clustering of contextual cascading bandits. They designed UCB-like algorithm CLUB-cascade to address the problem and provided an upper bound for its cumulative regret. Our work is based on Thompson sampling which tends to outperform UCB-like algorithms empirically [18]. We also give an alternative proof of the convergence rate of online clustering and provide a theoretical analysis of Thompson sampling in the contextual cascading settings.

3 Preliminaries

3.1 Problem Settings

We first formulate the collaborative contextual combinatorial cascading problem. In this problem, there are u users and these users can be partitioned into m clusters where $n \gg m$. The clusters are fixed but unknown to the learning agent. All users in the same cluster share the same preference which is encoded by a user vector $\theta \in \mathbb{R}^d$. For any users i and j, if they are not in the same cluster, then $\|\theta_i - \theta_j\| \geq \gamma$.

At each round t, the learning agent interacts with user i_t to customize personal recommendation. It first selects an ordered list of items $X_t = (X_{t,1}, X_{t,2}, ..., X_{t,k})$ from item set $\mathcal{X} \subset \mathbb{R}^d$ to recommend. The user checks the recommended list sequentially, clicks the first satisfying item and stops checking further. The learning agent observes the index of the clicked item C_t. It reveals that the first $C_t - 1$ items are not satisfying, the payoff of the C_t-th recommended item is 1, and the rest items are not checked by the user. If no item is clicked, the observed payoff will be $C_t = \infty$. The observed payoff $r(x)$ of an item x is generated by sampling from a Bernoulli distribution with

mean $\mathbb{E}[r(x)]$. The expected reward $\mathbb{E}[r(x)]$ of an item x is calculated by a linear function $\mathbb{E}[r(x)] = x^T\theta$. We assume that the probability of the user clicking each item is independent. Thus, the expected reward of a list X is

$$\mathbb{E}[r(X)] = 1 - \prod_{x \in X}(1 - x^T\theta).$$

It is worth noting that rearrangement of the items does not change the expected reward of a list. We define the optimal item list X^* as the list with maximum expected reward $X^* = \arg\max_{X \subset \mathcal{X}} \mathbb{E}[r(X)]$.

The instantaneous expected regret $R(t)$ at round t is defined as the gap between the expected reward of the optimal item list and that of the recommended list. The objective of the algorithm is to minimize the expected cumulative regret in T rounds:

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}[\sum_{t=1}^{T}(r(X^*) - r(X_t))],$$

where the expectation is taken over the randomness in selecting the recommended list X_t and the noise of the feedbacks.

3.2 Notations

We use $\|x\|_p$ to denote the p-norm of $x \in \mathbb{R}^d$. For matrix $M \in \mathbb{R}^{d \times d}$ and vector $x \in \mathbb{R}^d$, we denote by $\|x\|_M = \sqrt{x^T M x}$ the weighted 2-norm. We use $\lambda_{min}(M)$ and $\lambda_{max}(M)$ to denote the smallest and the largest eigenvalue of matrix M respectively.

Assumption 1 (Contextual vector and user vector). The contextual vectors and the user vector are in a closed subset of $\mathbb{R}^d$ such that $0 < ||x||_2^2 \leq 1$ for all $x \in \mathcal{X}$ and θ. This assumption is required so that the regret bound does not depend on the scale of the vectors. If $0 < ||x||_2^2 \leq L$, the regret bound would increase by a factor L.

Assumption 2 (Eigenvalues). For any round t, there exists a constant λ_{min} such that $\forall t$, $\lambda_{min} \leq \lambda_{min}(\mathbb{E}(x_t x_t^T))$. In standard contextual bandit algorithms, this assumption is often violated. The probability of selecting the optimal item will be 1 after enough rounds. Thus the smallest eigenvalue will be $\lambda_{min}(\mathbb{E}[x_t x_t^T]) = \lambda_{min}(x^* x^{*T}) = 0$. But in cascading settings, if the expected reward of the optimal item is smaller than 1, then the suboptimal items will be checked by the user with at least constant probability, thus $\lambda_{min}(\mathbb{E}[x_t x_t^T]) > \lambda_{min}$ is a reasonable assumption.

3.3 Collaborative Contextual Combinatorial Cascading Thompson Sampling

Our algorithm maintains a posterior distribution $\mathcal{N}(\hat{\theta}_t, V_t)$ of user vector θ for each user. The posterior distribution is updated with the recommended lists and the feedbacks. Let $(X_1, X_2, ..., X_n)$ be the sequence of lists recommended to one user and

$(C_1, C_2, ..., C_n)$ be the observed rewards until round t , the posterior distribution of user vector θ at round $t+1$ is $\mathcal{N}(\hat{\theta}_t, V_t^{-1})$, where $K_i = \min(K, C_i)$ and

$$V_t = \sum_{i=1}^{n}\sum_{k=1}^{K_i} X_{i,k}X_{i,k}^T, \quad f_t = \sum_{i=1}^{n}\sum_{k=1}^{K_i} X_{i,k}^T \mathbb{I}\{k = C_i\}, \quad \hat{\theta}_t = (\lambda I + V_t)^{-1} f_t. \quad (1)$$

Our algorithm also maintains an undirected graph $G_t(|u|, E_t)$ to store the cluster information. The graph is initialized as a fully-connected graph and the edges are removed gradually. At each round t, the algorithm first selects a user i_t to serve. It then finds the connected component set of i_t in graph G_{t-1}, which is referred to as M_t. The posterior distribution is then calculated by all the checked items and the feedbacks of the cluster M_t. The algorithm then samples $\tilde{\theta}_{i_t,M_t}$ from the distribution and generates the list by $X_{t,k} = \max_{x \in \mathcal{X} \setminus \{X_{t,1}, X_{t,2}, ... X_{t,k-1}\}} x^T \tilde{\theta}_{i_t,M_t}$. The algorithm then observes the feedback of the user and updates the user vector and the graph respectively.

Theorem 1. *For the collaborative contextual combinatorial cascading bandit problem, under Assumptions 1 and 2, the expected regret bound for C^4-TS algorithm within time horizon T is*

$$\mathbb{E}[R(T)] = O(d\sqrt{mKT}\ln KT + ud\ln duT)$$

Algorithm 1. C^4-Thompson sampling

Input: Set of items $\mathcal{X}$, $\lambda, \alpha, \beta > 0$
Init: $G_0 = (|u|, E_0)$ is a fully-connected graph over the user set $|u|$, for any user $i \in [u]$, $f_{i,0} = 0_d$, $V_{i,0} = \lambda I_d$, $\hat{\theta}_{i,0} = (V_{i,0} + \lambda I)^{-1} f_{i,0}$.
for $t = 1, 2, 3..., T$ **do**
  Select user i_t to serve and find the user set $M_t \subset |u|$ from graph G_{t-1} so that all users in M_t are connected to i_t
  Compute the following variable:
    $V_{i_t,M_t} = \lambda I + \sum_{j \in M_t} V_{j,t-1}$
    $f_{i_t,M_t} = \sum_{j \in M_t} f_{j,t-1}$
    $\hat{\theta}_{i_t,M_t} = V_{i_t,M_t}^{-1} f_{i_t,M_t}$
  Sample $\tilde{\theta}_{i_t,t}$ from distribution $\mathcal{N}(\hat{\theta}_{i_t,M_t}, \alpha V_{i_t,M_t}^{-1})$
  for $k \in [K]$ **do**
    Extract $X_{t,k} = \arg\max_{x \in \mathcal{X} \setminus \{X_{t,1}, X_{t,2}, ... X_{t,k-1}\}} x^T \tilde{\theta}_{i_t,t}$
  end for
  Recommend list X_t to user i_t and observe payoff C_t
  Set $r_t = \mathbb{I}(C_t \leq K)$ and $C_t = \min(C_t, K)$
  Update $f_{i_t,t}$, $V_{i_t,t}$, $\hat{\theta}_{i_t,t}$ and $N_{i_t,t}$ as in Equation (1)
  for $l \in [u]$ **do**
    if $\|\hat{\theta}_{i_t,t} - \hat{\theta}_{l,t}\|_2 \geq \beta(\frac{\sqrt{1+\ln(1+N_{i_t,t})}}{1+N_{i_t,t}} + \frac{\sqrt{1+\ln(1+N_{l,t})}}{1+N_{l,t}})$ **then**
      Delete the edges $(i_t, l) \in E_{t-1}$
    end if
  end for
end for

4 Regret Analysis

4.1 Proof Outline

The proof of Theorem 1 can be split into two parts. In the first part, we bound the expected number of rounds the algorithm need to partition the users into the right clusters, which is $O(ud \ln duT)$. In the second part, we prove that when the users are correctly partitioned, the expected regret bound is $O(d\sqrt{mKT} \ln KT)$. Thus the total regret is $\mathbb{E}[R(T)] = O(d\sqrt{mKT} \ln KT + ud \ln duT)$.

We follow three steps to show that the algorithm needs at most $O(ud \ln duT)$ rounds to partition the users into the right clusters. First, we notice that for any user, the 2-norm distance between $\hat{\theta}_t$ and θ decreases very fast [13]:

$$\|\hat{\theta}_t - \theta\|_2^2 \leq \frac{\|\hat{\theta}_t - \theta\|_{V_{t-1}}^2}{\lambda_{min}(V_{t-1})} = O(\frac{d \ln N_{t-1}}{N_{t-1}}),$$

where $\|\hat{\theta}_t - \theta\|_{V_{t-1}^{-1}}^2$ is the weighted 2-norm and $\lambda_{min}(V_{t-1})$ is the smallest eigenvalue of matrix V_{t-1}. Second, we prove that under Assumption 2, the smallest eigenvalue of the cumulative matrix V_{t-1} grows linearly with the number of checked arms $N_{t-1} = \sum_{i=1}^{t-1} C_i$ with high probability. Third, we model C_t as a truncated Poisson variable and show that after serving the user for $O(d \ln udT)$ rounds, the confidence interval for user vector will be smaller than $\gamma/2$, where the γ is the constant in the assumption of clusters. Thus, after $O(ud \ln udT)$ rounds, the edges between different clusters will be removed.

After the clusters are correctly partitioned, the recommendation is based on the estimates of cluster vector and its covariance matrix. We follow three steps to bound the expected cumulative regret up to round T. First, we define event F_k = {the k-th item in X_t is examined} for any time t and $k \in [K]$, and decompose the regret as [18]:

$$\mathbb{E}[R(t)] \leq \mathbb{E}[\sum_{k=1}^{K} \mathbb{I}(F_k)(r(X_k^*) - r(X_{t,k}))]$$

Thus, the instantaneous regret can be bounded by the difference between expected rewards of the best items and the checked items. Second, We define event $E^{\mu}(t)$, $E^{\theta}(t)$ and prove that these events happen with high probability. If both events are true, we further decompose the regret as

$$\mathbb{E}[R(t)] \leq \mathbb{E}[\sum_{k=1}^{C_t} (\alpha_t + \beta_t)(s_t(X_k^*) + s_t(X_{t,k}))],$$

where $s_t(x) = \|x\|_{V_{t-1}}$.

Finally, we show that under assumption 2, the smallest eigenvalue of the matrix V_{t-1} grows linearly with the number of items the user has observed, which is referred to as N_{t-1} in the algorithm. We can then prove that the sum of the variance $\sum_{t=1}^{T} \sum_{k=1}^{C_t} s_t(X_k^*)$ is of order $\sqrt{dKT \ln KT}$. Then, substituting this result along with Lemma 2, we obtain the desired expected regret bound:

$$\mathbb{E}[R(T)] = O(d\sqrt{mKT} \ln KT + ud \ln duT)$$

4.2 Proof of Part 1

Definition 1. *Define $\mathbb{E}^{\lambda}(t)$ as the event that the smallest eigenvalue of V_t grows linearly with N_t, where N_t is the number of checked items after t rounds. Formally, define $\mathbb{E}^{\lambda}(t)$ as the event that*

$$\lambda_{min}(V_t) \geq 1/2N_t \cdot \lambda_{min}, \quad \forall N_t \geq (\frac{8}{\lambda_{min}^2} + \frac{4}{3\lambda_{min}}) \ln \frac{dut^2}{\delta},$$

We prove that event $\mathbb{E}^{\lambda}(t)$ holds with probability at least $1 - \frac{\delta}{ut^2}$ by substituting $\delta = \delta'/ut^2$ into Lemma 6.

Definition 2. *Define $\mathbb{E}^{\theta}(t)$ as the events that the estimator $\hat{\theta}$ is close to its real value θ. More precisely, define $\mathbb{E}^{\theta}(t)$ as the event that*

$$|\hat{\theta}_{t+1} - \theta|_{V_t} \leq \alpha_t,$$

where $\alpha_t(\delta) = R\sqrt{d \ln \frac{ut^2(1+N_t/\lambda)}{\delta}} + \sqrt{\lambda}$.

We prove that event $\mathbb{E}^{\theta}(t)$ holds with probability at least $1 - \frac{\delta}{ut^2}$ by substituting $\delta = \delta'/ut^2$ into Lemma 5.

If the events $\mathbb{E}^{\lambda}(t)$ and $\mathbb{E}^{\theta}(t)$ both hold for all users, then for any user,

$$\|\hat{\theta}_{t+1} - \theta\|_2 \leq \frac{\|\hat{\theta}_{t+1} - \theta\|_{V_t}}{\sqrt{\lambda_{min}(V_t)}} \leq \frac{\sqrt{2}\alpha_t}{\sqrt{N_t \lambda_{min}}} \leq \frac{\gamma}{2},$$

where the last inequality is valid when

$$N_t \geq \frac{8d}{\lambda_{min}\gamma^2} \ln \frac{uT^2}{\delta}.$$

Combining with the condition in $\mathbb{E}^{\lambda}(t)$, it is required that

$$N_t \geq \max\{\frac{8d}{\lambda_{min}\gamma^2} \ln \frac{uT^2}{\delta}, (\frac{8}{\lambda_{min}^2} + \frac{4}{3\lambda_{min}}) \ln \frac{duT^2}{\delta}\}.$$

If the user has been served in t rounds, where

$$t \geq \frac{2K^2}{q^2} \ln \frac{8duT}{\delta} + \frac{2}{q}\{\frac{8d}{\lambda_{min}\gamma^2} \ln \frac{uT^2}{\delta}, (\frac{8}{\lambda_{min}^2} + \frac{4}{3\lambda_{min}}) \ln \frac{duT^2}{\delta}\} := t_0,$$

the algorithm will be able to partition the user into the real cluster with probability at least $1 - 4\delta/u$. The above inequality is proven by modeling C_t as a truncated Poisson variable with mean q where $q = O(K/p)$. And the lower bound of t is calculated by using Lemma 4.

It reveals that the after ut_0 rounds, the user clusters are correctly partitioned with probability at least $1 - 4\delta$. Thus the cumulative regret before all the users are correctly partitioned is

$$\mathbb{E}[R'(T)] = O(ud \ln duT)$$

4.3 Proof of Part 2

After the users are correctly clustered, the information learned by a user is shared by all users in the same cluster. And the users in different clusters are independent. We consider the cumulative regret of one cluster. Suppose the users in the cluster are served in T rounds.

Following the previous approach [10,11], we rearrange the elements of the optimal list X^* so that if $x \in X^*$ and $x \in X_t$, then $index(X^*, x) = index(X_t, x)$. Under this arrangement, for all round t,

$$\forall k \in [K], \qquad X_k^{*T}\theta \geq X_{t,k}^T\theta \quad and \quad X_k^{*T}\tilde{\theta}_t \leq X_{t,k}^T\tilde{\theta}_t.$$

The algorithm uses the user feedbacks and contextual vector to update the estimator $\hat{\theta}_t$ and the covariance matrix V_t^{-1}. As the algorithm accumulates more information each round, $\hat{\theta}$ approaches θ gradually and the variance of expected reward of each item decreases. If $\tilde{\theta}$, $\hat{\theta}$, and θ are close enough, the algorithm is likely to select the optimal list and the regret can be bounded by the variance. This intuition leads to the definition of the following two events.

Definition 3. *Define $\mathbb{E}^\theta(t)$ and $\mathbb{E}^\mu(t)$ as the events that $x^T\hat{\theta}_t$ and $x^T\tilde{\theta}_t$ are concentrated around $x^T\theta$ and $x^T\hat{\theta}_t$ respectively. Formally, define $\mathbb{E}^\theta(t)$ and $\mathbb{E}^\mu(t)$ as*

$$Event \quad \mathbb{E}^\theta(t) : \forall x \in \mathcal{X} : |x^T\hat{\theta}_t - x^T\theta| \leq \alpha_t s_t(x)$$
$$Event \quad \mathbb{E}^\mu(t) : \forall x \in \mathcal{X} : |x^T\tilde{\theta}_t - x^T\hat{\theta}_t| \leq \beta_t s_t(x),$$

where $\alpha_t(\delta) = R\sqrt{d \ln \frac{t^2(1+N_t/\lambda)}{\delta}} + \sqrt{\lambda}$, $\beta_t = \sqrt{4d \ln \frac{t}{\delta}}$ *and* $s_t(x) = \|x\|_{V_{t-1}^{-1}}$.

We prove in Lemma 5 that both events hold with probability at least $1-\delta/t^2$. And if events $\mathbb{E}^\mu(t)$ and $\mathbb{E}^\theta(t)$ are both true, the instantaneous expected regret can be decomposed as:

$$\begin{aligned}
\mathbb{E}[R(t)] &= \mathbb{E}[\prod_{k\in[K]}(1 - r(X_k^*)) - \prod_{k\in[K]}(1 - r(X_{t,k}))] \\
&\leq \mathbb{E}[\sum_{k=1}^{K}[\prod_{j=1}^{k-1}(1 - r(X_{t,j}))](r(X_k^*) - r(X_{t,k}))] \\
&\leq \mathbb{E}[\sum_{k=1}^{K}\mathbb{I}(F_{t,k})(r(X_k^*) - r(X_{t,k}))] \\
&\leq \mathbb{E}[\sum_{k=1}^{K}\mathbb{I}(F_{t,k})(\alpha_t + \beta_t)(s_t(X_k^*) + s_t(X_{t,k}))], \qquad (2)
\end{aligned}$$

where $F_{t,k}$ is defined as the event that the item $X_{t,k}$ is checked by the user. Equation (2) is by

$$\begin{aligned}
r(X_k^*) - r(X_{t,k}) &\leq (X_k^{*T} - X_{t,k}^T)\tilde{\theta}_t + (\alpha_t + \beta_t)(\|X_k^*\|_{V_{t-1}^{-1}} + \|X_{t,k}\|_{V_{t-1}^{-1}}) \\
&\leq (\alpha_t + \beta_t)(s_t(X_k^*) + s_t(X_{t,k})).
\end{aligned}$$

If event $\mathbb{E}^{\lambda}(t)$ holds, then the expected cummulative regret is

$$\begin{aligned}\mathbb{E}[R(T)] &\leq \sum_{t=1}^{T}\mathbb{E}[\sum_{k=1}^{K}\mathbb{I}(F_{t,k})(\alpha_t+\beta_t)(s_t(X_k^*)+s_t(X_{t,k}))] \\ &\leq (\alpha_T+\beta_T)\sum_{t=1}^{T}\mathbb{E}[\sum_{k=1}^{C_t}s_t(X_k^*)+\mathbb{E}[\sum_{k=1}^{C_t}s_t(X_{t,k})] \\ &\leq (\alpha_T+\beta_T)(\sqrt{\frac{2d}{\lambda_{min}}}(2\sqrt{TK\ln TK}+K)+\sqrt{2dTK\ln(1+\frac{KT}{\lambda d})}).\end{aligned} \tag{3}$$

If event $\mathbb{E}^{\lambda}(t)$ is true for any round t, then $\lambda_{min}(V_{t-1}) \geq 1/2N_{t-1}\lambda_{min}$. This implys that for any item $x \in \mathcal{X}$, $\|x\|_{V_{t-1}^{-1}} \leq \sqrt{\frac{2d}{N_{t-1}\lambda_{min}}}$, where $N_{t-1} = \sum_{i=1}^{t-1} C_i$ is number of checked items. Thus applying Lemma 7, we get that

$$\begin{aligned}\sum_{t=1}^{T}\sum_{k=1}^{C_t}s_t(X_k^*) &\leq \sqrt{\frac{2d}{\lambda_{min}}}(C_1+\sum_{t=2}^{T}\frac{C_t}{\sqrt{\sum_{i=1}^{t-1}C_i}}) \\ &= \sqrt{\frac{2d}{\lambda_{min}}}(\sqrt{2KT\ln KT}+K).\end{aligned}$$

And the second term of Eq. (3) follows from Lemma 2.

Substituting the value of α_T and β_T in $\mathbb{E}[R(T)]$, we obtain that for one cluster, if the users in the cluster are served in T rounds, the cumulative regret is:

$$\mathbb{E}[R(T)] = O(d\sqrt{KT}\ln KT).$$

Suppose the users are partitioned into m clusters and each cluster is served in $T_1, T_2, ..., T_m$ rounds where $\sum_{i=1}^{m} T_i = T$, the total regret is

$$\begin{aligned}\mathbb{E}[R''(T)] &= \sum_{i=1}^{m}\mathbb{E}[R(T_i)] \\ &\leq d\ln KT\sum_{i=1}^{m}C_i\sqrt{KT_i} \\ &= O(d\sqrt{mKT}\ln KT)\end{aligned}$$

Combining the results of part 1 and part 2 completes the proof of Theorem 1.

4.4 Technique Lemmas

In this section, we introduce the technique lemmas used in the proof of Theorem 1.

Lemma 1. *(Confidence Ellipsoid [1]). Let $(x_t : t \geq 0)$ be a sequence of d-dimensional vectors and $\|x_t\|^2 \leq 1$. Let $r_t = x_t^T\theta + \epsilon_t$ where ϵ_t is R-sub-Gaussian for some constant*

R, $V_t = \lambda I + \sum_{i=1}^{t} x_t x_t^T$ *and* $\hat{\theta}_t = V_t^{-1} \sum_{i=1}^{t} x_i r_i$. *Then, for any* $0 < \delta < 1$ *and* $t \geq 1$,

$$||\hat{\theta}_t - \theta||_{V_t} \leq \alpha_t(\delta)$$

holds with probability at least $1 - \delta$, *where*

$$\alpha_t(\delta) = R\sqrt{d \ln \frac{1 + t/\lambda}{\delta}} + \sqrt{\lambda}.$$

Lemma 2. *(Sum of standard deviation [13]). Let* $\lambda > 1$, *for any sequence* $(X_1, X_2, ..., X_T)$, *let* $V_t = \lambda I + \sum_{i=1}^{t} \sum_{k=1}^{C_t} X_{i,k} X_{i,k}^T$ *where* $C_t \leq K$. *Then*

$$\sum_{t=1}^{T} \sum_{k=1}^{C_t} \|X_{t,k}\|_{V_{t-1}^{-1}} = O(\sqrt{dTK \ln(1 + \frac{TK}{\lambda d})}).$$

Lemma 3. *(Azuma-Hoeffding inequality [4,9]). If* $(Y_t : t \geq 0)$ *is a super-martingale process, and for all* $t \in [T]$, $|Y_{t+1} - Y_t| \leq c_t$ *for some constant* c_t, *then for any* $a \geq 0$,

$$P(Y_T - Y_0 \geq a) \leq 2e^{-\frac{a^2}{2\sum_{t=1}^{T} c_t^2}}.$$

Lemma 4. *(Sum of variables) Let* $(C_t : t \geq 1)$ *be a sequence of truncated Poisson variables with mean* $1 \leq q_i \leq K$ *and* $q = \min_{i \in [t]}\{q_i\}$. *Let* $\delta > 0$ *and* $B > 0$, *then*

$$\sum_{i=1}^{t} C_i \geq B$$

holds for all $t \geq \frac{2B}{q} + \frac{2}{q^2} k^2 \ln \frac{2}{\delta}$ *with probability at least* $1 - \delta$.

Proof. We construct a super-martingale process by defining $X_i = C_i - q_i$ and $Y_i = \sum_{j=1}^{i} X_j$. By the Azuma-Hoeffding inequality (Lemma 3), we obtain that for all $t \geq \frac{2B}{q} + \frac{2}{q^2} k^2 \ln \frac{2}{\delta}$,

$$P(\sum_{i=1}^{t} C_i \leq B) = P(\sum_{i=1}^{t} (q_i - C_i) \geq \sum_{i=1}^{t} q_i - B) \leq 2e^{-\frac{(tq-B)^2}{2tk^2}} \leq \delta.$$

Lemma 5. *(High probability property of the events). For all* t *and* $0 < \delta < 1$, *event* $\mathbb{E}^{\mu}(t)$ *happends with probability at least* $1 - \frac{\delta}{t^2}$. *And for any possible filtration event* $\mathbb{E}^{\theta}(t)$ *happens with probability at least* $1 - \frac{1}{t^2}$.

Proof. The proof of this lemma follows from previous work on linear Thompson sampling [3]. The high probability property of $\mathbb{E}^{\mu}(t)$ is proven by applying the concentration inequality stated as Lemma 7 in [1]. The probability bound for $\mathbb{E}^{\theta}(t)$ is obtained by applying the concentration inequality of Gaussian random variables [2].

Lemma 6. *(Lower bound of the smallest eigenvalue of sum of hermitian matrices). Let* $(x_t x_t^T : t \geq 1)$ *be a sequence of* $d \times d$ *matrices generated sequentially from random distribution* $x_t x_t^T \in \mathbb{R}^{d\times d}$*. Suppose that for all* t*,* $\mathbb{E}[x_t x_t^T]$ *is full rank Hermitian matrix and* $\mathbb{E}[x_t x_t^T] \geq \lambda_{min}$ *(Assumption 2) and* $\|x\| \leq 1$ *(Assumption 1). Let* $V_t = \sum_t x_t x_t^T$*, then for any* $t \geq (\frac{8}{\lambda_{min}^2} + \frac{4}{3\lambda_{min}}) \ln \frac{d}{\delta}$*, event* $\lambda_{min}(V_t) \geq 1/2t\lambda_{min}$ *holds with probability at least* $1 - \delta$*.*

Proof. We first define three random sequences:

$$X_t = \mathbb{E}[x_t x_t^T] - x_t x_t^T$$
$$Y_t = \sum_{k=1}^{t} X_k = \sum_{k=1}^{t} \mathbb{E}[x_k x_k^T] - \sum_{k=1}^{t} x_k x_k^T$$
$$W_t = \sum_{k=1}^{t} \mathbb{E}_{k-1}[X_k^2].$$

As $\mathbb{E}X_t = \mathbb{E}[\mathbb{E}[x_t x_t^T] - x_t x_t^T] = 0$, Y_t is a matrix martingale whose values are Hermitian matrices with dimension $d \times d$ and X_t is the difference sequence. Note that $\lambda_{max}(X_t) \leq 1$, then by the Matrix Freedman's inequality, for any a and b:

$$\mathbb{P}(\lambda_{max}(Y_t) \geq a \quad and \quad \lambda_{max}(W_t) \leq b) \leq d \cdot \exp^{-\frac{a^2/2}{b+a/3}}.$$

We define that $V_t = \sum_{k=1}^{t} x_k x_k^T$ and $G_t = \sum_{k=1}^{t} \mathbb{E}[x_k x_k^T]$, then $V_t + Y_t = G_t$. By the Wely's Theorem $\lambda_{min}(V_t) + \lambda_{max}(Y_t) \geq \lambda_{min}(G_t)$, then

$$\begin{aligned}
\mathbb{P}(\lambda_{min}(V_t) \geq \frac{1}{2}t\lambda_{min}) &\geq \mathbb{P}(\lambda_{min}(G_t) - \lambda_{max}(Y_t) \geq \frac{1}{2}t\lambda_{min}) \\
&= \mathbb{P}(\lambda_{max}(Y_t) \leq \lambda_{min}(G_t) - \frac{1}{2}t\lambda_{min}) \\
&= 1 - \mathbb{P}(\lambda_{max}(Y_t) \geq \lambda_{min}(G_t) - \frac{1}{2}t\lambda_{min}) \\
&\geq 1 - \mathbb{P}(\lambda_{max}(Y_t) \geq \frac{1}{2}t\lambda_{min}) \qquad (4) \\
&= 1 - \mathbb{P}(\lambda_{max}(Y_t) \geq \frac{1}{2}t\lambda_{min} \quad and \quad \|W_t\| \leq t) \qquad (5) \\
&\geq 1 - d \cdot \exp(-\frac{\lambda_{min}^2 t^2/8}{t + 1/6\lambda_{min} t}).
\end{aligned}$$

Equation (4) holds because of the assumption that $\lambda_{min}(\mathbb{E}[x_t x_t^T]) \geq \lambda_{min}$ and the Wely's inequality and Equation (5) holds because of the fact that $\lambda_{max}(W_t) = \lambda_{max}(\sum_{k=1}^{t} \mathbb{E}[(x_k x_k^T)^2] - \mathbb{E}[x_k x_k^T]^2) \leq t$ holds with probability 1.

Thus, for any $t \geq (\frac{8}{\lambda_{min}^2} + \frac{4}{3\lambda_{min}}) \ln \frac{d}{\delta}$, $\mathbb{P}(\lambda_{min}(V_t) \geq 1/2t\lambda_{min}) \geq 1 - \delta$, which completes the proof.

Lemma 7. *Suppose* $S = (a_t : t \in [T])$ *is a finite sequence of positive integer and* $1 \leq a_i \leq K$ *for any* $i \leq T$*. Let* $f(S) = a_1 + \sum_{t=2}^{T} \frac{a_t}{\sqrt{\sum_{j=1}^{t} a_j}}$*. Then,* $f(S) = O(K + \sqrt{KT \ln KT})$*.*

Proof. We first prove that for any sequence $S_1 = (a_1, a_2, ..., a_T)$, if there exist $1 \leq i \leq T$ that $a_i \geq a_{i+1}$, we can switch these two elements a_{i+1} and a_i so that we get another sequence $S_2 = (a_1, a_2, ..., a_{i+1}, a_i, ...a_T)$ and $f(S_2) \geq f(S_1)$. We set that $\sum_{j=1}^{i-1} a_j = M$, then

$$\begin{aligned} f(S_2) - f(S_1) &= \frac{a_{i+1}}{\sqrt{M}} + \frac{a_i}{\sqrt{M + a_{i+1}}} - \frac{a_i}{\sqrt{M}} - \frac{a_{i+1}}{\sqrt{M + a_i}} \\ &= \frac{a_{i+1} - a_i}{\sqrt{M}} + \frac{a_i}{\sqrt{M + a_{i+1}}} - \frac{a_{i+1}}{\sqrt{M + a_i}} \end{aligned}$$

We define a function $g(x) = \frac{1}{\sqrt{M+x}}$, as $g''(x) = \frac{3}{4}(M + x)^{-\frac{5}{2}} \geq 0$, $g(x)$ is a convex function. Then,

$$\forall x_1, x_2 \geq 0 \quad and \quad t \in [0, 1], \quad g(tx_1 + (1 - t)x_2) \leq tg(x_1) + (1 - t)g(x_2).$$

We substitute $x_1 = 0$, $x_2 = a_{i+1}$, $t = 1 - a_i/a_{i+1}$ into above inequality, and we obtain

$$g(a_i) \leq (1 - \frac{a_i}{a_{i+1}})g(0) + \frac{a_i}{a_{i+1}} g(a_{i+1})$$

Thus,

$$\frac{a_{i+1}}{\sqrt{M + a_i}} \leq \frac{a_{i+1} - a_i}{\sqrt{M}} + \frac{a_i}{\sqrt{M + a_{i+1}}},$$

so that $f(S_2) \geq f(S_1)$.

For any sequence S, if the elements of the sequence are fixed, we can recursively switch the elements to get the maximum value of $f(S)$. The maximal value is obtain when $a_1 \leq a_2 \leq ... \leq a_T$. As the value of the elements can only be selected from K positive integers, we assume that the integer $1 \leq k \leq K$ is selected T_k times and $\sum_{k=1}^{K} T_k = T$, thus

$$\begin{aligned} a_1 + \sum_{i=2}^{T} \frac{a_i}{\sqrt{\sum_{i=1}^{i} a_i}} &\leq K + \sqrt{(\sum_{i=1}^{T} a_i)(\sum_{i=2}^{T} \frac{a_i}{\sum_{j=1}^{i} a_j})} && (6) \\ &\leq K + \sqrt{KT(\sum_{i=2}^{T} \frac{a_i}{\sum_{j=1}^{i} a_j})} \\ &\leq K + \sqrt{KT(\sum_{k=1}^{K} \sum_{j=1}^{T_k} \frac{k}{\sum_{h=1}^{k-1} hT_h + (j-1)k})} && (7) \\ &\leq K + \sqrt{KT(\sum_{k=2}^{K} \ln \frac{k}{k-1} + \ln \sum_{k=1}^{K} \frac{k}{K} T_k)} && (8) \\ &= O(K + \sqrt{KT \ln KT}), \end{aligned}$$

where Eq. (6) follows from the Cauchy–Schwarz inequality and $a_i \leq K$, Eq. (7) follows from the fact that after the rearrangement of the sequence, $a_i \leq a_{i+1}$ holds for any $1 \leq i \leq T-1$, and $\sum_{k=1}^{K} T_k = T$. Equation (8) holds because $\sum_{j=1}^{T_k} \frac{k}{\sum_{h=1}^{k-1} hT_h + (j-1)k} \leq \ln \sum_{i=1}^{k} \frac{i}{k} T_i - \ln \sum_{i=1}^{k-1} \frac{i}{k} T_i$.

5 Experiment

We evaluate our algorithm C^4-TS on a synthetic dataset. Its performance is compared with CLUB-cascade, CascadeLinUCB and CascadeLinTS, which are the most related algorithms. The empirical results demonstrate the advantage of using Bayesian heuristic and online clustering.

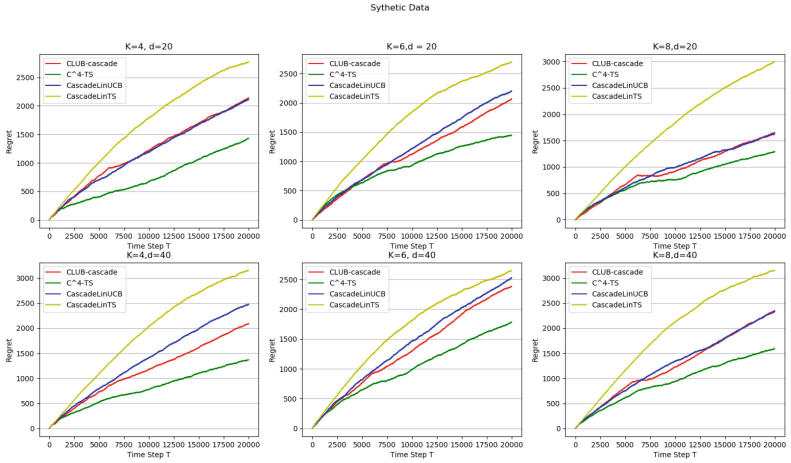

Fig. 1. These figures compare C^4-TS with CLUB-cascade, CascadeLinUCB and CascadeLinTS on Synthetic dataset. Plots reporting the cummulative regret over time step T. The basic setting is that there are 200 items, 20 users and 2 clusters. The users in different clusters have orthogonal user vectors. The dimension of feature space is $d = \{20, 40\}$. The length of recommended list is $K = \{4, 6, 8\}$

In all the subfigures, we generate a candidate set with $N_{items} = 200$ items, each item is represented by a d-dimensional feature vector $x \in \mathbb{R}^d$ with $x^T x \leq 1$ and $d \in \{20, 40\}$. We then generate $N_{users} = 20$ users and the users can be grouped into two clusters. We set that the users in the same cluster share the same user vectors $\theta \in \mathbb{R}^d$. And for users in different clusters, we set that their user vectors are orthogonal so that $\gamma = \sqrt{2}$ in these settings. The observed payoff for user θ to item x is a Bernoulli random variable, whose mean is the linear function $x^T\theta$. At each round, the algorithm selects a user to serve and recommends $K = \{4, 6, 8\}$ items to the user. The algorithm then observes the cascading feedback and updates its parameters accordingly.

In Fig. 1, we plot the cumulative regret as a function of time step T for C^4-TS, CLUB-cascade, CascadeLinUCB and CascadeLinTS. It is obvious that our algorithm

outperforms other algorithms in all settings. We compare the performance of the collaborative algorithms and the standard bandit algorithm. It can be seen that the collaborative algorithms significantly outperform those algorithms without online clustering in all settings, which demonstrates the advantage of utilizing collaborative effect in these algorithms. We can also compare the performance of Thompson sampling and UCB-like algorithms. Although CascadeLinTS does not perform as well as CascadeLinUCB, our algorithm outperforms CLUB-cascade in all settings. In fact, we can tune CascadeLinTS by adjusting the exploration rate so that CascadeLinTS performs as well as CascadeLinUCB, but Fig. 1 is a clear proof of the collaborative effect on Thompson sampling. It can be seen that the collaborative algorithms benefit from collaborative effects after several rounds. This observation is empirical evidence of part 1 that the algorithm can find the cluster structure efficiently. Another important property of Thompson sampling is that it often has higher variance than UCB. An explanation is that Thompson sampling requires additional randomness because it samples from the posterior distribution of θ to explore information. In contrast, UCB-based algorithms explore by adding a deterministic positive bias.

6 Conclusion

We design and analyze C^4-TS algorithm for the stochastic cascading bandit in a collaborative environment. We prove that the regret bound of our algorithm matches the regret bound for UCB-like algorithms. And the experiments conducted on a synthetic dataset demonstrate the advantage of our algorithm over existing UCB-like algorithms and standard Thompson sampling algorithm. Further investigations include deriving the lower regret bound for cascading bandit and the frequentist regret bound for Thompson sampling algorithms.

Acknowledgments. This paper is supported by the National Science Foundation of China under Grant 61472385 and Grant U1709217.

References

1. Abbasi-Yadkori, Y., Pál, D., Szepesvári, C.: Improved algorithms for linear stochastic bandits. In: Advances in Neural Information Processing Systems, pp. 2312–2320 (2011)
2. Abramowitz, M., Stegun, I.: Handbook of mathematical functions with formulas, graphs, and mathematical tables (applied mathematics series 55). National Bureau of Standards, Washington, DC (1964)
3. Agrawal, S., Goyal, N.: Thompson sampling for contextual bandits with linear payoffs. In: International Conference on Machine Learning, pp. 127–135 (2013)
4. Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Math. J. **19**(3), 357–367 (1967). Second Series
5. Chapelle, O., Li, L.: An empirical evaluation of Thompson sampling. In: Advances in Neural Information Processing Systems, pp. 2249–2257 (2011)
6. Cheung, W.C., Tan, V.Y.F., Zhong, Z.: Thompson sampling for cascading bandits (2018)
7. Gentile, C., Li, S., Kar, P., Karatzoglou, A., Etrue, E., Zappella, G.: On context-dependent clustering of bandits. arXiv preprint arXiv:1608.03544 (2016)

8. Gentile, C., Li, S., Zappella, G.: Online clustering of bandits. In: International Conference on Machine Learning, pp. 757–765 (2014)
9. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**(301), 13–30 (1963)
10. Kveton, B., Szepesv, C., Wen, Z., Ashkan, A.: Cascading bandits: learning to rank in the cascade model (2015)
11. Li, S., Karatzoglou, A., Gentile, C.: Collaborative filtering bandits. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 539–548. ACM (2016)
12. Li, S., Wang, B., Zhang, S., Chen, W.: Contextual combinatorial cascading bandits. In: Proceedings of the 33rd International Conference on Machine Learning, pp. 1245–1253 (2016)
13. Li, S., Zhang, S.: Online clustering of contextual cascading bandits (2018)
14. Nguyen, T.T., Lauw, H.W.: Dynamic clustering of contextual multi-armed bandits. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1959–1962. ACM (2014)
15. Russo, D., Van Roy, B.: Learning to optimize via posterior sampling. Math. Oper. Res. **39**(4), 1221–1243 (2014)
16. Russo, D., Van Roy, B.: An information-theoretic analysis of Thompson sampling. J. Mach. Learn. Res. **17**(1), 2442–2471 (2016)
17. Russo, D.J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al.: A tutorial on Thompson sampling. Found. Trends Mach. Learn. **11**(1), 1–96 (2018)
18. Shi, Z., Hao, N., Sung, K., Nan, R.K., Kveton, B.: Cascading bandits for large-scale recommendation problems. In: Conference on Uncertainty in Artificial Intelligence (2016)
19. Wu, Q., Wang, H., Gu, Q., Wang, H.: Contextual bandits in a collaborative environment. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 529–538. ACM (2016)