Contextual Combinatorial Conservative Bandits

Xiaojin Zhang,1 Shuai Li,2 Weiwen Liu,1 Shengyu Zhang 3
1The Chinese University of Hong Kong, 2Shanghai Jiao Tong University, 3Tencent
xjzhang@cse.cuhk.edu.hk, lishuai.sherry@gmail.com, wwliu@cse.cuhk.edu.hk, shengyuzhang@gmail.com

Abstract

The problem of multi-armed bandits (MAB) asks to make sequential decisions while balancing between exploitation and exploration, and have been successfully applied to a wide range of practical scenarios. Various algorithms have been designed to achieve a high reward in a long term. However, its short-term performance might be rather low, which is injurious in risk sensitive applications. Building on previous work of conservative bandits, we bring up a framework of contextual combinatorial conservative bandits. An algorithm is presented and a regret bound of $\tilde{O}(d^2 + d\sqrt{T})$ is proven, where $d$ is the dimension of the feature vectors, and $T$ is the total number of time steps. We further provide an algorithm as well as regret analysis for the case when the conservative reward is unknown. Experiments are conducted, and the results validate the effectiveness of our algorithm.

1 Introduction

The problem of multi-armed bandits has been extensively studied and drawn lot of attention in the past decades [6]. In canonical stochastic multi-armed bandit problem, the learner is presented with a set of arms, whose rewards are independently and identically distributed. The learner is allowed to select one arm at each round, and the final goal is to maximize the cumulative rewards. The key challenge of the bandit problem lies in the exploitation-exploration trade-off. On the one hand, the historical decisions and observed rewards could be exploited to find the arm with the highest empirical mean reward so far; on the other hand, the learner could explore other arms to get a better estimate of their mean value, which helps to achieve a large cumulative reward in a long run.

In some practical applications, the learning agent needs to select a set of arms at each round, instead of just one arm. For example, a company has newly hired several employees, and a certain amount of tasks are awaited to be completed. These employees could all deal with these tasks, but the completion qualities might be distinct. Actually, the rewards gained by the employees dealing with various tasks follow some specific distributions, while the company has no prior information about the distributions. The company could adjust the assignment policy based on the observations of the employees’ performance, in hope of allocating the tasks wisely to maximize the cumulative revenue. This kind of problem fits into the framework of combinatorial bandits. Each employee-task pair could be regarded as an arm, the allocation policy is an action, and the feasible action set contains allocations which could form a matching.

Contextual bandit problem is another extended version of the traditional multi-armed bandit problem. It is supposed that the agent could observe certain contexts before making decisions at each round. Besides, the learner is aware of a hypothesis class, which could map contexts to arms and assist the learner in finding the arm with the highest reward. Indeed, the contextual bandit are sometimes referred to as partial-label, associative bandit, multi-armed bandit with expert advice and associative reinforcement learning [5]. The stochastic contextual bandit has been widely applied in lots of areas including online advertisement selection and news recommendation [7, 30, 29, 27], automated vaccine design and sensor management [16], influence maximization [12, 26], and offline evaluator construction [21].

Most algorithms proposed for either contextual bandit or combinatorial bandit aim at pursuing high reward, but usually ignore the instantaneous safety guarantee. For some risk sensitive applications such as financial investment, each action should be rather cautious to avoid the possibility that the cumulative return is below certain threshold at some rounds. As for the widely used Upper Confidence
Bound (UCB) strategy, it follows the principle that more explorations are in need especially when information is scarce, thus the policy is prone to explore the arms which have not been sufficiently sampled. From this point of view, this strategy is very effective for not missing any arm that might later have excellent performance. However, it is highly likely to choose some arms with low expected rewards but high variances at an early stage of the algorithm, resulting in a great amount of cumulative loss at some point of time. Similar concerns were recently addressed in [28], in which they introduced the conservative bandit problem and gave an algorithm that satisfied the revenue constraint uniformly in time under traditional bandit setting. Successively, [15] considered safety control in stochastic contextual bandit and assumed that the reward of each arm is linear with its feature vector.

In this paper, we make the following contributions. (1) We generalize the revenue constraint to a more widely used contextual combinatorial setting, obtaining a novel framework of contextual combinatorial conservative bandits. At each round, a set of arms could be selected, with the number of arms not exceeding a specified number. The learner could only observe the rewards of the played arms, which is referred to as semi-bandit feedback. Viewing that the reward of a set of arms is not purely in the linear combination form in most application scenarios, the reward function is generalized to non-linear case, satisfying two mild assumptions. (2) We then propose two algorithms, one for the case when the expected reward of the conservative arm set is prescribed, and one for the case when it is unknown. Both algorithms aim at maximizing the expected reward in the long run, while simultaneously satisfying the revenue constraint for each instant round. Importantly, our algorithm is computationally more efficient as compared with CLUCB proposed by [15] in the contextual conservative setting. (3) We provide detailed theoretical analysis for both algorithms, showing that the safety could be well maintained even in complex environments. Specifically, we consider the case when the dimension of the contextual information is rather large. With well-designed lower bound of the revenue constraint in the contextual combinatorial setting, we provide novel proofs and obtain a tighter bound than that presented in [15] in terms of high-dimensional contextual information, making it more practical for high-dimensional contextual scenarios. (4) We conduct several experiments by numerical simulation, the results are consistent with our theoretical analysis and confirm the advantage of our algorithm in terms of maintaining the cumulative reward above the specified safety threshold.

2 Related Work

Multi-armed bandit was initially proposed by [24]. In the traditional multi-armed bandit problem, only one arm is allowed to be selected at each round. [3] firstly extended the setting and assumed that multiple arms could be selected. However, the number of arms that could be selected at each round is fixed and constant. [14] generalized this setting, allowing any constraint on weights with regard to the selected arms, then a line of work associated with combinatorial bandit [9, 8, 22, 17, 10] were developed successively.

For many application scenarios like Internet advertisement selection and multi-modal message generation, contextual bandit turns out to be more suitable than the traditional bandit [19]. Different from semi-bandit feedback in which the learner could only observe the reward of the selected arm, the features of all arms are available in the contextual setting, which could assist the learner in estimating the rewards of the arms that are not selected and thus improve the performance. Consequently, a line of research considers the contextual bandit problem, typically under linear realizability assumption [13, 20, 11], where the expectation of each arm’s expected reward is a linear function of the features. To make the setting more applicable to online recommendation, [23] developed a novel framework called contextual combinatorial bandits and proposed an effective algorithm based on the upper confidence bound strategy.

In the traditional multi-armed bandits, the upper confidence bound strategy is not only intuitively sensible but also has been proven to be efficient and asymptotically optimal [18, 2, 4]. This strategy tends to follow the principle of optimism in face of uncertainty and ignores the risk it might face from a pessimistic point of view. However, safety insurance is the first priority in some areas including those related to health and finance. Recently, [28] paid attention to safety guarantee in the process of pursuing high rewards in the traditional multi-armed bandit setting, and proposed a conservative algorithm based on the upper confidence bound strategy. The algorithm maintains the constraint that the cumulative reward must not be less than a specific percentage of the reward gained by a conservative arm uniformly over time. Subsequently, [15] considered the revenue constraint in contextual bandits, under the assumption that the reward is linear and proposed an algorithm called conservative linear UCB (abbreviated as CLUCB). In order to calculate the confidence bounds of expected reward of each arm, it needs to search within a confidence set containing an infinite number of elements, while our algorithm could provide the bounds directly, leading to a great improvement in terms of the efficiency of the algorithm.
3 Problem Formulation

In the basic contextual combinatorial setting, the learner could select a subset of arms from a set \( E = \{1, 2, \ldots, M\} \) of \( M \) base arms at each round, also referred to as a super arm, subject to certain constraints on the selected arms. Suppose the maximum number of chosen items at each round is \( K \), then the set of feasible actions is the set of super arms with size less than or equal to \( K \), denoted by \( \Theta^K \). At round \( t \), each arm \( e \in E \) is associated with a feature vector \( x_{t,e} \in \mathbb{R}^d \), and the weight \( w_{t,e} \) could be represented as \( w_{t,e} = \theta^T x_{t,e} + \epsilon_{t,e} \), where \( \theta_e \in \mathbb{R}^d \) is a fixed but unknown parameter, and \( \epsilon_{t,e} \) is a random noise with zero mean. We take a standard assumption about the bound of length that \( \|\theta_e\|_2 \leq S \) and \( \|x_{t,e}\|_2^2 \leq L \), for all \( t \) and all \( e \in E \). Besides, the random noise is assumed to be conditionally 1-sub-Gaussian, i.e., for any \( \gamma \in \mathbb{R} \),

\[
E[\exp(\gamma \epsilon_{t,e})|\{A_{1:t-1}, w_{1:t-1}, x_{1:t}\}] \leq \exp(\gamma^2 / 2). \tag{1}
\]

Since the expectation of the random noise is zero, the expectation of the weight \( w_{t,e} \) is \( \theta^T x_{t,e} \), represented as \( w_{t,e}^* \).

After choosing a super arm \( A_t \) at round \( t \), the learning agent observes the weights of the arms in \( A_t \), and the expected reward of \( A_t \) is \( f(A, w_{t}^*) = f(w_{t}^*|A) \), where \( w|A = (w_e)_{e \in A} \) is a \( |A| \)-dimensional vector, \( w_{t}^* = (w_{t,e}^*)_{e \in E} \), and \( f \) is a function satisfying the following two natural properties.

- **Monotonicity** The function \( f(w) \) is non-decreasing with respect to \( w \), i.e., \( f(w) \leq f(w') \) if \( w_e \leq w_e' \) for all \( e \).

- **Lipschitz Continuity** For any two vectors \( w \) and \( w' \), we have \( |f(w) - f(w')| \leq P\|w - w'\|_2 \).

When it comes to the conservative setting, there is also a super arm \( A_0 \) whose expected weight is also given by \( w_{t,e}^* = \theta^T x_{t,e} \). In terms of the conservative perspective of the learner, the cumulative reward should not be less than a certain fraction of the reward gained by simply choosing the conservative arm set \( A_0 \) at each round, referred to as the revenue constraint:

\[
\sum_{t=1}^{T} f(A_s, w_{s,t}^*) \geq (1 - \alpha) f(A_0, w_{0,t}^*) \quad \forall t \in [T]. \tag{2}
\]

Here \( A_0 \) is the action selected by the conservative policy that our algorithm aims to out-perform, and its expected reward is \( \mu_0 := f(A_0, w_{0,t}^*) \). The parameter \( \alpha \) determines how conservative the agent should be, where \( \alpha \in (0, 1) \). Denote by \( A_t^* = \arg\max_{A \in \Theta^K} f(A, w_{t}^*) \), where \( \Theta^K = A_0 \cup \{A \subseteq E : |A| \leq K\} \) is the set of feasible actions. We assume that \( f(A, w_{t}^*) \geq 0 \) for any action \( A \in \Theta^K \), and there exist \( \Delta_{\text{min}} \) and \( \Delta_{\text{max}} \) such that \( \Delta_{\text{min}} \leq f(A_t^*, w_{t}^*) - \mu_0 \leq \Delta_{\text{max}} \) and \( \alpha \mu_0 + \Delta_{\text{min}} > 0 \).

The goal of the learner is to minimize the pseudo-regret, which is defined as

\[
R(T) = \sum_{t=1}^{T} [f(A_t^*, w_{t}^*) - f(A_t, w_{t}^*)]. \tag{3}
\]

4 Algorithms

We propose two algorithms based on the upper confidence bound strategy, to solve the contextual combinatorial bandit problem with the revenue constraint, in both cases when the conservative reward is prescribed and when it is unknown. Note that for the second case, the conservative arm set \( A_0 \) is prescribed while its expected reward is unknown.

4.1 Learning with Conservative Reward Known

The contextual combinatorial conservative bandits aimed at solving two key issues. One is to maximize the cumulative reward, the other is to guarantee the conservative constraint. For the first issue, we use the effective UCB approach, which selects the arms based on the upper confidence bound of the unknown expected weight. For the second issue, since the expected weight \( w_{t}^* \) in the LHS of the conservative constraint Eq.(2) is unknown, the learner constructs a lower confidence bound \( \psi_t \) of the LHS of Eq.(2). If

\[
\psi_t \geq (1 - \alpha)\mu_0, \tag{4}
\]

then the conservative constraint Eq.(2) is satisfied with high probability.

We now introduce a self-adaptive algorithm for the contextual combinatorial conservative bandit problem, assuming that enough statistics have been collected to get a good estimate of the expected reward \( \mu_0 \) of the conservative arm set. The pseudo-code of \textit{CCComUCB with Conservative Reward} is displayed in Algorithm 1, which could be divided into three main parts. First, we calculate the confidence intervals of the expected weight of each non-default arm (lines 5-18). Note that the upper and lower confidence bounds of the conservative arm set are both equal to the prescribed value \( \mu_0 \) (line 4). Second, define
Algorithm 1 CCCunUCB with Conservative Reward

1: **Input:** $\alpha, \Theta^K, \mu_0$
2: **Initialization:** $\theta_0 \leftarrow 0_{d \times 1}, \psi_0 \leftarrow \lambda I, Y_0 \leftarrow 0_{d \times 1}$, $N_0 = D_0 = \emptyset$, and $H_0 \leftarrow \sqrt{\lambda S} + \sqrt{\log(1/\delta^2)}$
3: for $t \leftarrow 1, 2, \ldots, T$ do
4: \[ f(A_0, U_{t,t}) \leftarrow \mu_0, f(A_t, L_{t,t}) \leftarrow \mu_0 \]
5: for $e \leftarrow 1, 2, \ldots, M$ do
6: \[ U_{t,e} \leftarrow \theta_{t-1}^T x_{t,e} + H_{t-1} \| x_{t,e} \|_{V_{t-1}^{-1}} \]
7: \[ L_{t,e} \leftarrow \max \{ 0, \theta_{t-1}^T x_{t,e} - H_{t-1} \| x_{t,e} \|_{V_{t-1}^{-1}} \} \]
8: end for
9: if $f(A_0, U_{t,t}) > \max_{A \in \Theta^K \setminus A_0} f(A, U_{t,t})$ then
10: \[ B_t = A_0 \]
11: else
12: \[ B_t = \arg \max_{A \in \Theta^K \setminus A_0} f(A, U_{t,t}) \]
13: end if
14: for $n \in N_{t-1}$ and $e \in A_n$ do
15: \[ L_{t,n,e} \leftarrow \max \{ 0, \theta_{t-1}^T x_{n,e} - H_{t-1} \| x_{n,e} \|_{V_{t-1}^{-1}} \} \]
16: end for
17: $\psi_t \leftarrow \sum_{n \in N_{t-1}} f(A_n, L_{t,n}) + f(B_t, L_{t,t}) + |D_{t-1}| \mu_0$
18: if $\psi_t \geq (1 - \alpha) \mu_0$ then
19: \[ A_t \leftarrow B_t, N_t \leftarrow N_{t-1} \cup \{ t \} \]
20: Observe $w_{t,e}$ for all $e \in A_t$
21: \[ Y_t \leftarrow V_{t-1}^{-1} + \sum_{e \in A_t} x_{t,e} x_{t,e}^T \]
22: \[ Y_t \leftarrow Y_{t-1}^{-1} + \sum_{e \in A_t} w_{t,e} x_{t,e} \]
23: \[ \theta_t \leftarrow V_{t-1}^{-1} Y_t \]
24: \[ H_t \leftarrow \sqrt{\lambda S} + \sqrt{\log (\det(V_t) / \lambda d^2)} \]
25: end if
26: $A_t \leftarrow A_0, t \leftarrow D_{t-1} \cup \{ t \}$
27: end if
28: end for

the best action $B_t$ to be either the default action $A_0$ or the non-default best action for the upper confidence bound of the expected weights, whichever gives the larger $f$ value. Third, depending on whether the revenue constraint is satisfied at this moment or not (lines 18 and 25), we use $B_t$ or $A_0$ (lines 19 and 26), respectively. However, as we do not know $w_{t,e}$ in Eq. (2), we do not precisely know whether the constraint is satisfied. Therefore we take a conservative option here: we compute a lower confidence bound $\psi_t$ on the cumulative reward $\sum_{s=1}^t f(A_s, w_{s,e}^*)$, and use $B_t$ if this lower confidence bound $\psi_t$ is already greater than or equal to the right hand side (RHS) of Eq. (2). If the optimistic arm set is selected, we also update some statistics based on the newly received contextual information and the observed weights (lines 19 and 24): the set $N_t$ (the set of rounds $s \leq t$ in which we use optimistic action), $D_t$ (the set of rounds $s \leq t$ in which we use the default action), the estimate $\hat{\theta}_t$ to $\theta_s$, and the confidence radius $H_t$ of the expected weight.

4.2 Learning with Conservative Reward Unknown

For new applications that suffer from a cold start, neither sufficient data nor profound experiences are available to provide a reliable estimation of the performance $\mu_0$ of the conservative policy. We modify Algorithm 1 to make it capable of handling the situation when the conservative reward is unknown, as shown in Algorithm 2 (Appendix A). The differences between these two algorithms lie in two aspects. On the one hand, since no reliable estimation of $\mu_0$ is available, we need to calculate the confidence bounds of the expected weight of each conservative arm (lines 18 and 27). On the other hand, to ensure that revenue constraint Eq. (2) is satisfied, we use a lower bound of the LHS in Eq. (2) and an upper bound of the RHS in Eq. (2) for comparison.

5 Regret Analysis

In order to maximize the cumulative reward, it is required to have a good grasp of the expected weight $w_{t,e}^*$ given the contextual information $x_{t,e}$. We are aware that the expected reward is linear with the contextual vector $x_{t,e}$, but the linear coefficient $\theta_t$ is unknown. In both algorithms, the coefficient vector is estimated using ridge regression solution $\hat{\theta}_t = V_{t-1}^{-1} Y_t$, where $V_t = \lambda I + X_t X_t^T, Y_t = X_t W_t$. Particularly, $X_t \in \mathbb{R}^{d \times \sum_{s \in N_t} |A_n|}$ has columns $x_{n,e}$ where $e \in A_n, n \in N_t$, and $W_t \in \mathbb{R}^{\sum_{s \in N_t} |A_n|}$ has rows $w_{n,e}$.

Denote by $L_{t,n,e} = \hat{\theta}_{t-1}^T x_{n,e} - H_{t-1} \| x_{n,e} \|_{V_{t-1}^{-1}}$ and $U_{t,n,e} = \hat{\theta}_{t-1}^T x_{n,e} + H_{t-1} \| x_{n,e} \|_{V_{t-1}^{-1}}$. Then, according to the following lemma, $L_{t,n,e}$ is a lower bound of the expected weight $w_{t,e}^*$, and $U_{t,n,e}$ is an upper bound of $w_{t,e}^*$, where $e \in E$.

**Lemma 1.** Assume that $\| \theta_s \|_2 \leq S$ and $\| x_{t,e} \|_2 \leq L$, for all $t$ and all $e \in E$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have that for all $t \geq 1$ and $s \leq t$,

\[ \hat{\theta}_{t-1}^T x_{s,e} - H_{t-1} \| x_{s,e} \|_{V_{t-1}^{-1}} \]

and

\[ \hat{\theta}_{t-1}^T x_{s,e} + H_{t-1} \| x_{s,e} \|_{V_{t-1}^{-1}} \]

where the radius $H_{t-1}$ is

\[ H_{t-1} = \sqrt{\lambda S} + \sqrt{\log (\det(V_{t-1}) / \lambda d^2)} \].

The proof of Lemma 1 is illustrated in Appendix B. Note that Lemma 1 displays the confidence interval of the expected reward with the instantly updated variables instead of the static estimation at the specific round. Specifically,
with the contextual information $x_{s,e}$ at round $s$, Lemma 1 informs us the confidence interval of the expected reward based on the updated $\hat{\theta}_{t-1}$, $V_{t-1}$ and $H_{t-1}$ at any subsequent round $t$ ($t \geq s$).

In the following analysis, we use $d_t = |D_t|$ and $n_t = |N_t|$ to represent the total number of using the conservative policy and that of the optimistic policy separately.

**Lemma 2.** For any $t \geq 1$, $\text{det}(V_t) \leq (\lambda + n_t KL/d)^d$.

We proceed the proof of Lemma 2 by taking advantage of the relationship between the trace and the determinant, as shown in Appendix C.

It follows from Lemma 2 that

$$H_t = \sqrt{\lambda S} + \sqrt{\log(\text{det}(V_t)/\lambda^d)}$$

$$\leq \sqrt{\lambda S} + 2\log(1/\delta) + d\log(1 + KLt/\lambda d).$$

Denote by $C_t = \sqrt{\lambda S} + 2\log(1/\delta) + d\log(1 + KLt/\lambda d) + \sqrt{\lambda S}$, we have $L_{t,s,e} \geq \theta_{t-1}^T x_{s,e} - C_{t-1}\|x_{s,e}\|_{V_{t-1}}$, and $U_{t,s,e} \leq \theta_{t-1}^T x_{s,e} + C_t\|x_{s,e}\|_{V_{t-1}}$.

The cumulative regret of choosing the optimistic policies can be bounded as presented in the following lemma.

**Lemma 3.** Assume that $\|\theta_s\|_2 \leq S$ and $\|x_{t,e}\|_2 \leq L$, for any $t \geq 1$ and any $e \in A_t$. If $\lambda \geq L$, then for any $\delta$, with probability at least $1 - \delta$ and for any $T \geq 1$, the regret bound for selecting the optimistic policy is

$$\sum_{t \in T_t} [f(A_t, U_t) - f(A_t, w_t)]$$

$$\leq 2PT \sqrt{2dT \log \left(1 + \frac{n_T KL}{\lambda d}\right)},$$

where $C_T = \sqrt{2\log(1/\delta) + d\log(1 + T KL/\lambda d)} + \sqrt{\lambda S}$.

Note that we have extended the result to the case when $\|x_{t,e}\|_2^2 \leq L$ and at most $K$ arms are allowed to be selected at each round. When $n_T = T$, the regret bound corresponds to that of the standard contextual combinatorial bandits, and the order of which is $O(d\sqrt{T} \max\{1, \log (TK/d)\})$. The detailed proof procedure is presented in Appendix D.

The following two lemmas present two upper bounds on $\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_2^2$. The first bound is smaller when $n_t K \leq d$, while the second bound is smaller when $n_t K > d$.

**Lemma 4.** For any $t \geq 1$, suppose that $\|x_{t,e}\|_2^2 \leq L$, then

$$\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_2^2 \leq n_t K \left[1 - \frac{\lambda}{\lambda + n_t KL/d}\right].$$

**Proof.** By the definition of $V_t$, it can be represented as $V_t = \lambda I + X_t X_t^T$, where $X_t \in \mathbb{R}^{d \times n_t}$. The columns are $x_{n,a_n}, k \in [A_n], n \in N_t$. Thus,

$$\text{det}(\lambda I) = \text{det}(V_t - X_t X_t^T)$$

$$= \text{det}(V_t - V_t^{-1} X_t X_t^T)$$

$$= \text{det}(V_t) \cdot \text{det}(I - V_t^{-1} X_t X_t^T)$$

$$= \text{det}(V_t) \cdot \text{det}(I - X_t^T V_t^{-1} X_t),$$

where the last equality follows from the fact that $\text{det}(I - AB) = \text{det}(I - BA)$.

Note that the diagonal elements of $X_t^T V_t^{-1} X_t$ are $\|x_{n,a_n}\|_{V_t^{-1}}^2, k \in [A_n], n \in N_t$. Thus, we have

$$\text{trace}(I - X_t^T V_t^{-1} X_t) = \sum_{n \in N_t} |A_n|$$

$$- \sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_2^2.$$  \hspace{1cm} (9)

Denote by $N = \sum_{n \in N_t} |A_n|$, then $I - X_t^T V_t^{-1} X_t \in \mathbb{R}^{N \times N}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the eigenvalues of $I - X_t^T V_t^{-1} X_t$, then

$$\text{det}(I - X_t^T V_t^{-1} X_t) = \lambda_1 \times \lambda_2 \times \ldots \lambda_N$$

$$\leq (\lambda_1 + \lambda_2 + \ldots + \lambda_N)^N / N^N$$

$$= (\text{trace}(I - X_t^T V_t^{-1} X_t) / N)^N.$$  \hspace{1cm} (10)

By (8), (9), and (10), we obtain

$$\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_2^2 \leq N \left[1 - (\text{det}(\lambda I)/\text{det}(V_t))^{1/N}\right],$$

combined with Lemma 2 gives

$$\text{det}(V_t) \leq (\lambda + n_t KL/d)^d.$$  \hspace{1cm} (11)

Thus,

$$\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_2^2 \leq N \left[1 - \left(\frac{\lambda}{\lambda + n_t KL/d}\right)^{d/N}\right].$$

Denote by $a = \lambda / (\lambda + n_t KL/d)$, $f(x) = x(1 - a^{d/x})$, then

$$f'(x) = 1 - a^{d/x} (1 + d/x \ln(1/a)).$$  \hspace{1cm} (11)
Since \( u > \ln(u) + 1 \) for \( u > 1 \), we have \( f'(x) > 0 \). Thus
\[
\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 
\leq n_t K \left[ 1 - \left( \frac{\lambda}{\lambda + n_t KL/d} \right)^{d/(n_t K)} \right].
\]
\( \square \)

Next we present a tighter bound for \( \sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 \)
when \( n_t K > d \), as stated in Lemma 5.

**Lemma 5.** For any \( t \geq 1 \), suppose that \( \|x_{t,e}\|_{2}^2 \leq L \), then
\[
\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 \leq \frac{n_t K L d}{\lambda d + n_t K L}.
\]

**Proof.**
\[
\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 
= \sum_{n \in N_t} \sum_{e \in A_n} \|x_{e}^T V_t^{-1/2} x_{e}\|_{2}^2 
= \text{trace} \left[ \sum_{n \in N_t} \sum_{e \in A_n} \sum_{n \in N_t} \sum_{e \in A_n} V_t^{-1/2} x_{e} x_{e}^T V_t^{-1/2} \right] 
= \text{trace} \left( V_t^{-1/2} (V_t - \lambda I) V_t^{-1/2} \right) 
= \text{trace} (I - \lambda V_t^{-1}),
\]

where (15) is due to the definition of \( V_t \) as \( V_t = \lambda I + \sum_{n \in N_t} \sum_{e \in A_n} x_{e} x_{e}^T \). Let \( \lambda_1, \lambda_2, \ldots, \lambda_d \) be the eigenvalues of \( V_t \), then \( 1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_d \) are the eigenvalues of \( V_t^{-1} \), and
\[
\text{trace}(V_t^{-1}) = \frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_d} 
\geq \frac{d^2}{\lambda_1 + \cdots + \lambda_d} 
\geq \frac{d^2}{d \lambda + n_t KL}.
\]

Thus,
\[
\sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 = d - \lambda \text{trace}(V_t^{-1}) 
\leq \frac{n_t K L d}{\lambda d + n_t KL}.
\]
\( \square \)

In terms of the construction of the lower bound of the expected weight \( w_{n,e}^* \) (line 15 in Algorithm 1 and line 18 in Algorithm 2), if we use \( L(n, n) \) instead of \( L(t, n) \), then the upper bound of \( d_t \) would depend on \( \sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 \) instead of \( \sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 \).

Importantly, the newly proposed upper bound of \( \sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 \) is tighter than that of \( \sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 \) proposed in [23].

**Lemma 6.** For any \( t \in D_T \), we have
\[
d_t < \left[ 1 - (1 + n_t) \alpha \right] \mu_0 - n_t \Delta_{\text{min}} / (\alpha \mu_0) 
+ 2 PC_t \sqrt{\frac{n_t \sum_{n \in N_t} \sum_{e \in A_n} \|x_{n,e}\|_{V_t^{-1}}^2 / (\alpha \mu_0)}},
\]

where \( C_t = \sqrt{\lambda S + 2 \log(1/\delta)} + d \log \left( 1 + \frac{K L t}{\lambda d} \right) \).

Lemma 6 provides an upper bound on the total time steps the default arm set is selected, and the detailed proof is illustrated in Appendix E.

It could be inferred from Lemma 1 that Algorithm 1 satisfies the conservative constraint Eq. (2) for all \( T \geq 1 \) with probability at least \( 1 - \delta \). Similarly, Algorithm 2 also ensures that the constraint holds for all \( T \geq 1 \) with probability at least \( 1 - \delta \). The following theorems illustrate the regret bounds for the proposed algorithms.

**Theorem 1.** If \( \lambda \geq L \), then the following regret bound is satisfied with probability at least \( 1 - \delta \)
\[
R(T) = O \left( d(d + \sqrt{T}) \max \{1, \log(\frac{T K}{d})\} + d \right).
\]

Furthermore, for \( TK \leq d \), we have
\[
R(T) = O \left( d \sqrt{T} + d^{3/2}/\sqrt{K} \right).
\]

Compared with the regret bound of order \( O(d\sqrt{T} \max \{1, \log(TK/d)\}) \) for the contextual combinatorial bandits as proposed in Lemma 3, the upper bound of CCCConUCB is added with an extra order \( O(d^2 \max \{1, \log(TK/d)\}) \), which is caused by the requirement of satisfying the revenue constraint \( (2) \) at each round. The detailed proof of Theorem 1 is shown in Appendix F.

**Theorem 2.** If \( \lambda \geq L \), then the following regret bound is satisfied with probability at least \( 1 - \delta \)
\[
R(T) = O(d\sqrt{T} \max \{1, \log(TK/d)\} 
+ \max \{d^2 \max \{1, \log(TK/d)\}, 
\sqrt{dK \max \{1, \log(TK/d)\}}\}).
\]
Table 1: The compliance of the revenue constraints of Contextual Combinatorial UCB and the alterations of the selection preference of CCConUCB against distinct $\alpha$.

| $\alpha$ | Contextual Combinatorial UCB | CCConUCB |
|---------|-----------------------------|-----------|
|         | # violated constraints | # satisfied constraints | # optimal policies | # conservative policies |
| 0.01    | 49,993                      | 7         | 519           | 49,481                 |
| 0.15    | 18,502                      | 31,498    | 46,698        | 3302                    |
| 0.3     | 10,666                      | 39,334    | 49,332        | 668                     |
| 0.6     | 2642                        | 47,358    | 49,752        | 248                     |
| 0.9     | 1032                        | 48,968    | 49,946        | 54                      |

where $C_T = \sqrt{2 \log \left( \frac{1}{\delta} \right)} + d \log \left( 1 + \frac{TKL}{\lambda d} \right) + \sqrt{\lambda S}$. Furthermore, if $TK \leq d$, then
\[
R(T) = O(d\sqrt{T} + d^{3/2}/\sqrt{K} + \sqrt{Kd}).
\] (23)

The proof of Theorem 2 is amenable to the technique used for the analysis of the CCConUCB when the conservative reward is available and is displayed in Appendix G.

6 Numerical Simulations

In this section, we conduct three simulation experiments. We compare the performance of distinct algorithms in various cases, record the frequency the non-conservative algorithm disobeys the conservative constraint, and analyze the alterations of the conservative learner’s preference between the conservative policy and the optimistic policy, and compare our conservative algorithm with a naive algorithm that always follows the conservative strategy.

In the first experiment, we simulated three UCB-based algorithms, which are the CCConUCB with Conservative Reward, CCConUCB without Conservative Reward, and $C^2$UCB (Contextual Combinatorial UCB) proposed in [23]. The result is displayed in Figure 1. A total of 100 arms are presented, each arm is associated with a 10-dimensional feature vector, and at most 2 arms are allowed to be selected at each round. For each arm, the elements of the feature vector are randomly sampled from the uniform distribution $U(-1, 1)$, satisfying that the expected reward of each arm is non-negative. The coefficient $\theta_\ast$ is randomly sampled from the normal distribution $N(0, I_{10})$. The conservative coefficient is $\alpha = 0.2$ for CCConUCB in both cases. For the conservative arms, the feature vector of each conservative arm is randomly sampled from the uniform distribution $U(-1, 1)$, satisfying that the expected reward of each arm is no less than the ninth best reward and can be up to the eighth best reward. The random noise is randomly sampled from the normal distribution $N(0, 1)$. As can be seen from the comparison between CCConUCB and Contextual Combinatorial UCB, a conservative learner might give up some opportunities for exploring more arms in order to maintain safety at each round, at the expense of a higher regret in the long run. In terms of the contextual combinatorial conservative setting, the algorithm with the information of the default reward performs better in terms of the average regret.
In the second simulation experiment, we record the total time steps Contextual Combinatorial UCB disregues the conservative rule in Table 1. The settings of the number of base arms, the maximum number of selected arms and the dimension of the contextual vector are identical with those in the first experiment. The smaller \( \alpha \) is, the more stringent the constraint would be, leading to more violations with regard to the non-conservative algorithm. We further analyze the selection preference of \( CConUCB \) between optimistic policy and conservative policy for distinct conservative coefficients. Viewing from the last two columns of Table 1, a more conservative learner tends to have a greater appetite for the conservative policy. With fewer chances for exploration, the agent learns slower about the non-conservative arms, resulting in a lower degrading speed in terms of the average expected regret, as shown in Figure 2. Consequently, although the regret of the conservative algorithm degrades at a slower speed with smaller \( \alpha \), it successfully avoids a large number of constraint violations.

In the third simulation experiment, we compared the regret of the \( CConUCB \) with a conservative policy, where the conservative arm set is always selected at each round. We record the endurance time the \( CConUCB \) needs to outperform the conservative policy, the result is shown in Figure 3. Since the gap between the optimal action and the conservative action is a random sample, we divide them into three groups and take average over each group, namely low, medium and high. The result indicates that the smaller the gap is, the longer it takes for our algorithm to outperform the conservative policy. Note that the cumulative reward of \( CConUCB \) is not less than \((1-\alpha)\) fraction of that of the conservative policy, and the endurance time depends also on \( \alpha \). The larger \( \alpha \) is, the less stringent the requirement is. With more chances to explore, the algorithm converges at a faster speed, then it would be much easier for the \( CConUCB \) algorithm to outperform the pure conservative algorithm.

7 Discussion and Conclusion

We introduce a novel framework of multi-armed bandits, referred to as contextual combinatorial conservative bandits. We propose two UCB-based algorithms dealing with two cases when the conservative reward is prescribed and unknown separately, which not only maximize the expected reward in the long run, but also ensure safety at any instantaneous round. The reward \( f(A, w) \) of the arm set \( A \) could be as simple as a summation of the reward of each arm \( a \in A \), and it can also be more complicated such as a sigmoid function. Besides, we remark that a tighter bound for high-dimensional contextual scenario is provided in the above theorem. Specifically, when introduced to the conservative contextual setting \( K = 1 \), and assume that \( T \leq d \), \textit{Algorithm 1} achieves a regret of \( O(d\sqrt{T} + d^{3/2}) \), which is tight than the order \( O(d\sqrt{T} \log(T) + d \log d^2) \) obtained by [15]. The baseline payoff could be time-variant or stationary, depending on the conservative learner and the specific application scenario. Our work focuses on the stationary case, i.e. the baseline payoff of the conservative learner does not change with time, while the time-variant case could be a considerable direction for future work.

References

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

[2] Rajeev Agrawal. Sample mean based index policies by \( o(\log n) \) regret for the multi-armed bandit problem. Advances in Applied Probability, 27(04):1054–1078, 1995.

[3] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: iid rewards. IEEE Transactions on Automatic Control, 32(11):968–976, 1987.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2-3):235–256, 2002.

[5] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandit algorithms with supervised learning guarantees. In AISTATS, pages 19–26, 2011.

[6] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed
bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[7] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

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

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

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

[11] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15, pages 208–214, 2011.

[12] Meng Fang and Dacheng Tao. Networked bandits with disjoint linear payoffs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1106–1115. ACM, 2014.

[13] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[14] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. 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.

[15] Abbas Kazerouni, Mohammad Ghavamzadeh, and Benjamin Van Roy. Conservative contextual linear bandits. *arXiv preprint arXiv:1611.06426*, 2016.

[16] Andreas Krause and Cheng S Ong. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2011.

[17] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*, 2015.

[18] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[19] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

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

[21] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

[22] Tian Lin, Bruno D Abrahao, Robert D Kleinberg, John Lui, and Wei Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *ICML*, pages 901–909, 2014.

[23] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. *SIAM*, pages 461–469, 2014.

[24] Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

[25] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.

[26] Sharan Vaswani, Branislav Kveton, Zheng Wen, Mohammad Ghavamzadeh, Laks Lakshmanan, and Mark Schmidt. Diffusion independent semi-bandit influence maximization. *arXiv preprint arXiv:1703.00557*, 2017.

[27] Yingfei Wang, Hua Ouyang, Chu Wang, Jianhui Chen, Tsvetan Asamov, and Yi Chang. Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *AAAI*, pages 2746–2753, 2017.

[28] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1254–1262, 2016.
[29] Baosheng Yu, Meng Fang, and Dacheng Tao. Linear submodular bandits with a knapsack constraint. In *AAAI*, pages 1380–1386, 2016.

[30] Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. In *Advances in Neural Information Processing Systems*, pages 2483–2491, 2011.
Algorithm 2 CCCConUCB without Conservative Reward

Algorithm 2

1: Input: $\alpha, \Theta^K$
2: Initialization: $\theta_0 \leftarrow 0_{d \times 1}, V_0 \leftarrow \lambda I, Y_0 \leftarrow 0_{d \times 1}, N_0 = D_0 = 0$, and $H_0 \leftarrow \sqrt{\lambda S} + \sqrt{\log(1/\delta^2)}$
3: for $t \leftarrow 1, 2, ..., T$ do
4: for $e \in A_0$ do
5: $U_{t,e} \leftarrow \hat{\theta}_{t-1}^T x_{0,e} + H_{t-1} \| x_{0,e} \|_{V_{t-1}^{-1}}$
6: $L_{t,e} \leftarrow \max \{ 0, \hat{\theta}_{t-1}^T x_{0,e} - H_{t-1} \| x_{0,e} \|_{V_{t-1}^{-1}} \}$
7: end for
8: for $e \leftarrow 1, 2, ..., M$ do
9: $U_{t,e} \leftarrow \hat{\theta}_{t-1}^T x_{t,e} + H_{t-1} \| x_{t,e} \|_{V_{t-1}^{-1}}$
10: $L_{t,e} \leftarrow \max \{ 0, \hat{\theta}_{t-1}^T x_{t,e} - H_{t-1} \| x_{t,e} \|_{V_{t-1}^{-1}} \}$
11: end for
12: if $f(A_0, U_{t,0}) > \max_{A \in \Theta^K \backslash A_0} f(A, U_{t,t})$ then
13: $B_t = A_0$
14: else
15: $B_t = \arg \max_{A \in \Theta^K \backslash A_0} f(A, U_{t,t})$
16: end if
17: for $n \in N_{t-1}$ and $e \in A_n$ do
18: $L_{t,n,e} \leftarrow \max \{ 0, \hat{\theta}_{t-1}^T x_{n,e} - H_{t-1} \| x_{n,e} \|_{V_{t-1}^{-1}} \}$
19: end for
20: if $\sum_{n \in N_{t-1}} f(A_n, L_{t,n}) + f(B_t, L_{t,t}) + |D_{t-1}| \cdot (1 - \alpha)f(A_0, U_{t,0})$ then
21: $A_t = B_t, N_t \leftarrow N_{t-1} \cup \{ t \}$
22: Observe $w_{t,e}$ for all $e \in A_t$
23: $V_t \leftarrow V_{t-1} + \sum_{e \in A_t} x_{t,e} x_{t,e}^T$
24: $Y_t \leftarrow Y_{t-1} + \sum_{e \in A_t} y_{t,e} y_{t,e}^T$
25: $\hat{\theta}_t \leftarrow V_t^{-1} Y_t$
26: $H_t \leftarrow \sqrt{\lambda S} + \sqrt{\log \left( \frac{\det(V_t)}{\lambda^d \delta^2} \right)}$
27: else
28: $A_t \leftarrow A_0, D_t \leftarrow D_{t-1} \cup \{ t \}$
29: end if
30: end for

Proof. \[
| (\theta_s - \hat{\theta}_{t-1})^T x_{s,e} | = | (\theta_s - \hat{\theta}_{t-1})^T V_{t-1}^{1/2} V_{t-1}^{-1/2} x_{s,e} |
\leq \| (\theta_s - \hat{\theta}_{t-1})^T V_{t-1}^{1/2} \|_2 \cdot \| x_{s,e} V_{t-1}^{-1/2} \|_2
= \| (\theta_s - \hat{\theta}_{t-1})^T \|_{V_{t-1}^{-1}} \cdot \| x_{s,e} \|_{V_{t-1}^{-1}}
\leq H_{t-1} \| x_{s,e} \|_{V_{t-1}^{-1}},
\]
where the second inequality follows from Theorem 2 in \cite{1} (For the convenience of the reader, also recalled as Lemma 1.1 in the Appendix). \hfill \Box

Lemma 1. \cite{1} Assume that $\| \theta_s \|_2 \leq S$, then for any $\delta > 0$, with probability at least $1 - \delta$ for all $t \geq 0$, we have
\[
\| (\theta_s - \hat{\theta}_t)^T \|_{V_t} \leq \sqrt{\lambda S} + \sqrt{\log \left( \frac{\det(V_t)}{\lambda^d \delta^2} \right)}.
\]

C Proof of Lemma 2

Lemma 2. For any $t \geq 1$, $\det(V_t) \leq (\lambda + n_t KL/d)^d$.

Proof. Denote the eigenvalues of $V_t$ as $\lambda_1, \lambda_2, \ldots, \lambda_d$, then
\[
\det(V_t) = \lambda_1 \times \lambda_2 \times \ldots \times \lambda_d \\
\leq \left( \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_d}{d} \right)^d \\
= \left( \frac{\max(\lambda_1, \lambda_2, \ldots, \lambda_d)}{d} \right)^d.
\]
Since $V_t$ can be represented as $V_t = \lambda I + \sum_{s \in N_t, e \in A_s} x_{s,e} x_{s,e}^T$, it gives that
\[
\text{trace}(V_t) = \text{trace}(\lambda I) + \sum_{s \in N_t} \sum_{e \in A_s} \text{trace}(x_{s,e} x_{s,e}^T)
= \lambda d + \sum_{s \in N_t} \sum_{e \in A_s} \| x_{s,e} \|_2^2
\leq \lambda d + n_t KL.
\]
Combining (25) and (26), we obtain
\[
\det(V_t) \leq (\lambda + n_t KL/d)^d.
\]
\hfill \Box

D Proof of Lemma 3

Lemma 3. Assume that $\| \theta_s \|_2 \leq S$ and $\| x_{t,e} \|_2^2 \leq L$, for any $t \geq 1$ and all $e \in A_s$. If $\lambda \geq L$, then for any $\delta$, with
\[
H_{t-1} = \sqrt{\lambda S} + \sqrt{\log \left( \frac{\det(V_{t-1})}{\lambda^d \delta^2} \right)}.
\]
probability at least $1 - \delta$ and for any $T \geq 1$, the regret bound for selecting the optimistic policy is

$$\sum_{t \in N_T} [f(A_t, U_t) - f(A_t, w^*_t)] \leq 2PC_T \sqrt{2dn_T \log \left(1 + \frac{n_T KL}{d\lambda}\right)},$$

where

$$C_T = \frac{\sqrt{\lambda S} + \sqrt{2} \log \left(1 + \frac{1}{\delta}\right) + d \log \left(1 + \frac{T KL}{d\lambda}\right)}{\sqrt{2}}.$$

**Proof.** According to the property claimed in Lemma 1, we know that $U_t \geq w^*_t$, combined with the monotonicity of the reward function, we have $f(A^*_t, w^*_t) \leq f(A^*_t, U_t)$. Since $A_t = \arg\max_{A_t \in A} f(\lambda, A_t)$, we also have $f(A^*_t, U_t) \leq f(A_t, U_t)$. The cumulative regret of selecting the optimistic policy can be bounded as follows,

$$\sum_{t \in N_T} [f(A_t, U_t) - f(A_t, w^*_t)] \leq P \sum_{t \in N_T} \sqrt{\sum_{e \in A_t} 4C_T^2 \|x_{t,e}\|^2_{v_{t-1}^{-1}}} \leq P \sqrt{\sum_{t \in N_T} \sum_{e \in A_t} 4n_T C_T^2 \|x_{t,e}\|^2_{v_{t-1}^{-1}}} \leq 2PC_T \sqrt{\sum_{t \in N_T} \sum_{e \in A_t} n_T \|x_{t,e}\|^2_{v_{t-1}^{-1}}}.$$

If $\lambda \geq L$, then $\|x_{t,e}\|^2_{v_{t-1}^{-1}} \leq \frac{\|x_{t,e}\|^2_{v_{t-1}^{-1}}}{\lambda} \leq \frac{L}{\lambda} \leq 1$. Since $x \leq 2 \log(1 + x)$ for $x \in [0, 1]$, we have

$$\sum_{t \in N_T} \sum_{e \in A_t} \|x_{t,e}\|^2_{v_{t-1}^{-1}} \leq 2 \log \left(1 + \sum_{e \in A_t} \|x_{t,e}\|^2_{v_{t-1}^{-1}}\right).$$

Denote the rounds selecting the optimistic arm sets as $t_1, t_2, \ldots, t_{n_T}$. From the recursion formula of $V_t$, we have

$$\det(V_{t_{n_T}}) = \det(V_{t_{n_T-1}} + \sum_{e \in A_t} x_{t_{n_T}, e}(x_{t_{n_T}, e})^T) = \det(V_{t_{n_T-1}}) \cdot \det \left( I + \sum_{e \in A_t} (V_{t_{n_T-1}}^{-1/2} x_{t_{n_T}, e})(V_{t_{n_T-1}}^{-1/2})^T \right) \geq \det(V_{t_{n_T-1}}) \left( 1 + \sum_{e \in A_t} \|x_{t_{n_T}, e}\|^2_{T_{t_{n_T-1}}} \right)$$

In addition, Lemma 2 informs us that

$$\det(V_T) \leq (\lambda^d + n_T KL/d)^d.$$

Thus,

$$\sum_{t \in N_T} [f(A_t, U_t) - f(A_t, w^*_t)] \leq 2PC_T \sqrt{2dn_T \log \left(1 + \frac{n_T KL}{d\lambda}\right)}.$$

**E Proof of Lemma 6**

**Lemma 6.** For any $t \in D_T$, we have

$$d_t < (1 - (1 + n_t)\alpha) \mu_0 - n_t \Delta_{\text{min}})/(\alpha \mu_0) + 2PC_t \sqrt{\frac{n_t}{\sum_{n \in N_t} \sum_{e \in \mathcal{A}_n} \|x_{n,e}\|^2_{v_{t-1}}}/(\alpha \mu_0)},$$

where $C_t = \sqrt{\lambda S} + \sqrt{2} \log(1/\delta) + d \log \left(1 + \frac{KLt}{d\lambda}\right).$
Proof. Suppose the conservative arm set $A_0$ is selected at round $t$, then according to Algorithm 1, it is satisfied that
\[
\sum_{n \in N_{t-1}} f(A_n, L_{t,n}) + f(B_t, L_{t,t}) + d_{t-1}\mu_0 < (1 - \alpha)t\mu_0 \tag{34}
\]

Note that $t$ can be denoted as $t = n_{t-1} + d_{t-1} + 1$. By dropping $f(B_t, L_{t,t})$, and rearranging the terms in (34), we have
\[
\alpha d_{t-1}\mu_0 < [1 - (1 + n_{t-1})\alpha] \mu_0 + n_{t-1}\mu_0 - \sum_{n \in N_{t-1}} f(A_n, L_{t,n})
\]
\[
= [1 - (1 + n_{t-1})\alpha] \mu_0 + \sum_{n \in N_{t-1}} [\mu_0 - f(A_n, U_{t,n}) + f(A_n, U_{t,n}) - f(A_n, L_{t,n})]
\]
\[
\leq [1 - (1 + n_{t-1})\alpha] \mu_0 + \sum_{n \in N_{t-1}} [\mu_0 - f(A_n, U_{t,n}) + f(A_n, U_{t,n})]
\]
\[
+ P \left[ \sum_{e \in A_n} (U_{t,n,e} - L_{t,n,e})^2 \right] \tag{35}
\]
\[
= [1 - (1 + n_{t-1})\alpha] \mu_0 + \sum_{n \in N_{t-1}} [\mu_0 - f(A_n, U_{t,n}) + f(A_n, U_{t,n})]
\]
\[
+ 2PC_{t-1} \sqrt{\sum_{e \in A_n} \|x_{n,e}\|^2_{V^{-1}_{t-1}}} \tag{36}
\]
\[
\leq [1 - (1 + n_{t-1})\alpha] \mu_0 - n_{t-1}\Delta_{min}
\]
\[
+ 2PC_{t-1} \sqrt{n_{t-1} \sum_{n \in N_{t-1}} \sum_{e \in A_n} \|x_{n,e}\|^2_{V^{-1}_{t-1}}} \tag{37}
\]

where (35) is due to the lipschitz continuity of the reward function, (36) follows from Lemma 1, and (37) follows from Cauchy-Schwarz inequality.

**F Proof of Theorem 1**

**Theorem 1.** If $\lambda \geq L$, then the following regret bound is satisfied with probability at least $1 - \delta$
\[
R(T) < 2PC_T \sqrt{2dT \log (1 + TKL/(\lambda d))}
\]
\[
+ \left[ \frac{1}{\alpha} + (\alpha\mu_0 + \Delta_{min}) \lambda d/ (\alpha\mu_0 KL) + 4P^2 C_{T-1}^2 \sqrt[4]{3\alpha\mu_0 (\alpha\mu_0 + \Delta_{min})} \right] \Delta_{max}
\]
\[
= O \left( \frac{d(d + \sqrt{T}) \max\{1, \log(\frac{TK}{d})\} + \frac{d}{\sqrt{K}}}{T} \right),
\]

where $C_T = \sqrt{X_S} + \frac{2 \log(1/\delta) + d \log(1 + \frac{TKL}{\lambda d})}{\sqrt{2}}$, Furthermore, for $TK \leq d$, we have
\[
R(T) < 2PC_T \sqrt{2dT \log (1 + TKL/(\lambda d))}
\]
\[
+ \left[ \frac{1}{\alpha} + \frac{2Pd}{\alpha\mu_0} \sqrt{\frac{L}{K (L + \lambda C_{T-1})}} \right] \Delta_{max}
\]
\[
= O \left( \frac{d\sqrt{T} + d^{3/2}/\sqrt{K}}{} \right).
\]

Proof. Suppose $t'$ is the last round the conservative policy is played, then
\[
d_{t'-1} = d_T - 1, \tag{38}
\]
and by Lemma 6 we have
\[
\alpha d_{t'-1}\mu_0 < [1 - (1 + n_{t'-1})\alpha] \mu_0 - n_{t'-1}\Delta_{min}
\]
\[
+ 2PC_{t'-1} \sqrt{n_{t'-1} \sum_{n \in N_{t'-1}} \sum_{e \in A_n} \|x_{n,e}\|^2_{V^{-1}_{t'-1}}} \tag{39}
\]
\[
< [1 - (1 + n_{t'-1})\alpha] \mu_0 - n_{t'-1}\Delta_{min}
\]
\[
+ 2PC_{t'-1} n_{t'-1} \sqrt{\frac{KLd}{\lambda d + n_{t'-1}KL}},
\]
where (39) follows from Lemma 5.

Define
\[
f(x) = - (\alpha\mu_0 + \Delta_{min}) x + 2PC_{T-1}x \sqrt{\frac{KLd}{\lambda d + KLx}},
\]
de note $(\alpha\mu_0 + \Delta_{min})/KL = A$, $2PC_{T-1} \sqrt{\frac{d}{KL}} = B$, then
\[
f(x) = -AKLx + \frac{BKLex}{\sqrt{KLx + \lambda d}},
\]
with the substitution $u = KLx + \lambda d$, we have
\[
f(u) = -Au + A\lambda d + B\sqrt{u} - \frac{B\lambda d}{\sqrt{u}}. \tag{40}
\]

Note that
\[
f''(u) = -\frac{B}{4}u^{-3/2} - \frac{3B\lambda d}{4}u^{-5/2} < 0,
\]
which signifies that the global maximum of $f$ could be reached at the unique stationary point. Set the first order $f'(t) = 0$ gives
\[
\frac{B\lambda d}{\sqrt{u}} = 2At - B\sqrt{u}. \tag{41}
\]
Substitute (41) into (40) gives
\[
f(u) = -3At + 2B\sqrt{u} + A\lambda d \leq A\lambda d + \frac{B^2}{3A}.
\]
Thus, \( d_T = d'_{t-1} + 1 \)
\[
< \frac{1}{\alpha} + (\alpha \mu_0 + \Delta_{min}) \lambda d/ (\alpha \mu_0 K L) + \frac{4P^2 C^2_{t-1} d}{3(\alpha \mu_0 + \Delta_{min}) \alpha \mu_0}, \tag{42}
\]

According to the property claimed in Lemma 1, we know that \( U_t \geq w^*_t \), combined with the monotonicity of the reward function, we have \( f(A_t^*, w^*_t) \leq f(A_t^*, U_t) \). Since \( A_t = \text{argmax}_{A \in \Theta} f(A, U_t) \), we also have \( f(A_t^*, U_t) \leq f(A_t, U_t) \).

After running Algorithm 1 for \( T \) rounds, the cumulative regret can be bounded as follows,
\[
R(T) = \sum_{t=1}^{T} [f(A_t^*, w^*_t) - f(A_t, w^*_t)]
\[
= \sum_{t \in N_T} [f(A_t^*, w^*_t) - f(A_t, w^*_t)]
+ \sum_{t \in D_T} [f(A_t^*, w^*_t) - \mu_0]
\leq \sum_{t \in N_T} [f(A_t, U_t) - f(A_t, w^*_t)]
+ \sum_{t \in D_T} [f(A_t^*, w^*_t) - \mu_0]
\leq \sum_{t \in N_T} [f(A_t, U_t) - f(A_t, w^*_t)] + d_T \Delta_{max}
\tag{43}
\]

\[
< 2PC_T \sqrt{2dT \log \left( 1 + \frac{TKL}{d \lambda} \right)}
+ \left[ \frac{1}{\alpha} + (\alpha \mu_0 + \Delta_{min}) \lambda d/ (\alpha \mu_0 K L) + \frac{4P^2 C^2_{t-1} d}{3(\alpha \mu_0 + \Delta_{min}) \alpha \mu_0} \right] \Delta_{max}, \tag{45}
\]

where the first part of \((45)\) follows from Lemma 3 and the fact that \( n_T \leq T \), and the second part of \((45)\) follows from \((42)\).

Furthermore, if \( TK \leq d \), then it implies that \( n_{t'} K \leq d \), we apply Lemma 4 and have
\[
\alpha d_{t'-1} \mu_0
\[
< [1 - (1 + n_{t'-1} \alpha)] \mu_0 - n_{t'-1} \Delta_{min} + 2PC_{t'-1} n_{t'-1} \sqrt{K \left[ 1 - \left( \frac{\lambda}{\lambda + n_{t'-1} KL/d} \right)^{d/(n_{t'-1} K)} \right]}
\leq (1 - \alpha) \mu_0 + 2Pd \sqrt{\frac{L}{K (\lambda + L) C_{t'-1}}},
\]

Thus,
\[
d_T = d'_{t-1} + 1
\[
< \frac{1}{\alpha} + \frac{2Pd}{\alpha \mu_0} \sqrt{\frac{L}{K (\lambda + L) C_{t-1}}}. \tag{46}
\]

Combine \((46)\) with \((44)\) we have
\[
R(T) < 2PC_T \sqrt{2dT \log \left( 1 + \frac{TKL}{d \lambda} \right)}
+ \left[ \frac{1}{\alpha} + (\alpha \mu_0 + \Delta_{min}) \lambda d/ (\alpha \mu_0 K L) + \frac{4P^2 C^2_{t-1} d}{3(\alpha \mu_0 + \Delta_{min}) \alpha \mu_0} \right] \Delta_{max},
\]

\[
\leq O\left( d \sqrt{T} \max\{1, \log TK/d\} \right) + \max\{d^2 \max\{1, \log TK/d\},
\sqrt{d K \max\{1, \log TK/d\}}\},
\]

where \( C_T = \sqrt{2 \log \left( \frac{1}{\beta} \right) + d \log \left( 1 + \frac{TKL}{d \lambda} \right) + \sqrt{\lambda S}} \). Furthermore, if \( TK \leq d \), then
\[
R(T) < 2PC_T \sqrt{2dT \log \left( 1 + \frac{TKL}{d \lambda} \right)}
+ \left[ \frac{1}{\alpha} + 2P (1 - \alpha) \sqrt{KL/\lambda C_{t-1}} / (\alpha \mu_0) \right] \Delta_{max},
\]

\[
= O\left( d \sqrt{T} + d^{3/2} / \sqrt{K + \sqrt{Kd}} \right).
\]

Proof. If we want to choose the optimistic arm set at round \( t \), then the cumulative reward should not be less
than certain percent of the reward gained by choosing the conservative arm set. That is,
\[
\sum_{n \in N_{t-1}} f(A_n, w_n^*) + f(B_t, w_t^*) + \sum_{d \in D_{t-1}} f(A_0, w_0^*) - (1 - \alpha) \sum_{t=1}^{t} f(A_0, w_0^*) \tag{47}
\]
should be non-negative.

To make the selection at each round less risky, the conservative learner wants to ensure that the calculable lower bound of (47) is non-negative.

Consider that \( f(A_0, U_{t,0}) \geq f(A_0, w_0^*) \), then
\[
\sum_{n \in N_{t-1}} f(A_n, L_{t,n}) + f(B_t, L_{t,t}) + (d_{t-1} - (1 - \alpha) t) f(A_0, U_{t,0})
\]
is a lower bound of (47) when \( d_{t-1} < (1 - \alpha)t \). Suppose \( t' \) is the last round the conservative policy is played, then
\[
d_{t'} - 1 = d_T - 1, \tag{48}
\]
and
\[
\sum_{n \in N_{t'-1}} f(A_n, L_{t',n}) + f(B_{t'}, L_{t',t'}) + d_{t'-1} f(A_0, U_{t',0}) < (1 - \alpha) t' f(A_0, U_{t',0}). \tag{49}
\]
By Lemma 6, we have
\[
\alpha d_{t'-1} f(A_0, U_{t',0}) < [1 - (1 + n_{t'-1})\alpha] f(A_0, U_{t',0}) - n_{t'-1} \Delta_{min} + 2PC_{t'-1} \sqrt{n_{t'-1} \sum_{n \in N_{t'-1}} \sum_{e \in A_n} \|x_{n,e}\|_{V_{t'-1}}^2} \tag{50}
\]
From the lipschitz continuity of the expected reward function and Lemma 1, we have
\[
f(A_0, U_{t',0}) - f(A_0, w_0^*) \leq 2PC_{t'-1} \sqrt{\sum_{e \in A_0} \|x_{0,e}\|_{V_{t'-1}}^2} \tag{51}
\]
Since \( \|x_{0,e}\|_{V_{t'-1}} \leq \frac{\|x_{0,e}\|^2}{\sqrt{\lambda}} \leq \sqrt{\frac{L}{\lambda}} \), combined with (51), we have
\[
f(A_0, U_{t',0}) \leq f(A_0, w_0^*) + 2P \sqrt{\frac{\|A_0\| L}{\lambda} C_{t'-1}} \leq \mu_0 + 2P \sqrt{\frac{KL}{\lambda} C_{t'-1}}. \tag{52}
\]
If \( 1 - (1 + n_{t'-1})\alpha < 0 \), then from (50) and (52) we have
\[
\alpha d_{t'-1} \mu_0 < [1 - (1 + n_{t'-1})\alpha] \mu_0 - n_{t'-1} \Delta_{min} + 2PC_{t'-1} \sqrt{n_{t'-1} \sum_{n \in N_{t'-1}} \sum_{e \in A_n} \|x_{n,e}\|^2_{V_{t'-1}}}, \tag{53}
\]
which is consistent with the case when \( \mu_0 \) is known. Thus, we have
\[
d_T < 1/\alpha + (\alpha \mu_0 + \Delta_{min}) \lambda d / (\alpha \mu_0 KL) + 4P^2 C_{t'-1}^2 d \lambda / (\alpha \mu_0 KL) + 3(\alpha \mu_0 + \Delta_{min}) \alpha \mu_0 \Delta_{max}. \]
If \( 1 - (1 + n_{t'-1})\alpha \geq 0 \), then from (50) and (52) we have
\[
\alpha d_{t'-1} \mu_0 \leq (1 - \alpha) \left( \mu_0 + 2P \sqrt{\frac{KL}{\lambda} C_{t'-1}} \right) - n_{t'-1} \left( \mu_0 + \Delta_{min} + 2\alpha P \sqrt{\frac{KL}{\lambda} C_{t'-1}} \right) + 2PC_{t'-1} \sqrt{n_{t'-1} \sum_{n \in N_{t'-1}} \sum_{e \in A_n} \|x_{n,e}\|^2_{V_{t'-1}}}
\]
\[
\leq (1 - \alpha) \left( \mu_0 + 2P \sqrt{\frac{KL}{\lambda} C_{t'-1}} \right) - n_{t'-1} \left( \mu_0 + \Delta_{min} + 2\alpha P \sqrt{\frac{KL}{\lambda} C_{t'-1}} \right) + 2PC_{t'-1} n_{t'-1} \sqrt{\frac{KLd}{\lambda d + n_{t'-1} KL}}.
\]
where the second inequality follows from Lemma 5.

Denote by \( f(x) = -AKLx + \frac{BK Lx}{\sqrt{KLx} + \lambda d} \), where \( A = \frac{\alpha \mu_0 + \Delta_{min}}{\sqrt{KLx} + \lambda d} + \frac{2\alpha P}{\sqrt{KLx}} C_{t'-1}, B = 2PC_{t'-1} \sqrt{\frac{d}{KL}} \).
According to the proof of Theorem 1, we have \( f(x) \leq \).
Aλd + \frac{B^2}{3A}. \text{ Thus, }

\begin{align*}
d_T &= d_{t' - 1} + 1 \\
&< 1/\alpha + \frac{2(1 - \alpha)PC_{T-1}}{\alpha\mu_0} \sqrt{\frac{KL}{\lambda}} \\
&+ \left[ \frac{\alpha\mu_0 + \Delta_{\min}}{\alpha\mu_0 KL} + \frac{2\alpha P}{\alpha\mu_0 \sqrt{KL\lambda}} C_{T-1} \right] \lambda d \\
&+ \frac{4P^2C_{T-1}^2d}{3\left( \alpha\mu_0 + \Delta_{\min} + 2\alpha P \sqrt{KL/\lambda C_{T-1}} \right) \alpha\mu_0}.
\end{align*}

(54)

After running Algorithm 2 for \( T \) rounds, the cumulative regret could be bounded as follows,

\begin{align*}
R(T) &= \sum_{t=1}^{T} [f(A_t^*, w_t^*) - f(A_t, w_t^*)] \\
&\leq \sum_{t \in T_t} [f(A_t, U_t) - f(A_t, w_t^*)] + d_T \Delta_{\max}
\end{align*}

(55)

< 2PC_T \sqrt{2dT \log \left( 1 + \frac{TKL}{d\lambda} \right)} \\
+ \left[ \frac{1}{\alpha} + \frac{2(1 - \alpha)PC_{T-1}}{\alpha\mu_0} \sqrt{\frac{KL}{\lambda}} \right] \lambda d \\
+ \left[ \frac{\alpha\mu_0 + \Delta_{\min}}{\alpha\mu_0 KL} + \frac{2\alpha P}{\alpha\mu_0 \sqrt{KL\lambda}} C_{T-1} \right] \lambda d \\
+ \frac{4P^2C_{T-1}^2d}{3\left( \alpha\mu_0 + \Delta_{\min} + 2\alpha P \sqrt{KL/\lambda C_{T-1}} \right) \alpha\mu_0} \Delta_{\max}.
\end{align*}

(56)

where (55) follows from the proof in Theorem 1. (56) follows from Lemma 3 and (54).

Furthermore, if \( TK \leq d \), then it implies that \( n_{t'}K \leq d \).

If \( 1 - (1 + n'_{t'-1})\alpha < 0 \), the analysis is consistent with the case when \( \mu_0 \) is known. Thus,

\begin{align*}
d_T &= d_{t' - 1} + 1 \\
&< 1/\alpha + \frac{2Pd}{\alpha\mu_0} \sqrt{\frac{L}{K(\lambda + L) C_{T-1}}}.
\end{align*}

(57)

If \( 1 - (1 + n'_{t'-1})\alpha \geq 0 \), we apply Lemma 4 and have

\begin{align*}
\alpha d_{t' - 1} &\leq (1 - \alpha) \left( \mu_0 + 2P \sqrt{\frac{KL}{\lambda} C_{T-1}} \right) \\
&- n_{t' - 1} \left( \alpha\mu_0 + \Delta_{\min} + 2\alpha P \sqrt{KL/\lambda C_{T-1}} \right) \\
&+ 2PC_{T-1}n'_{t'-1}1 \sqrt{K \left[ 1 - \frac{\lambda}{\lambda + n_{t'-1}KL/d} \right]} \\
&\leq (1 - \alpha) \left( \mu_0 + 2P \sqrt{\frac{KL}{\lambda} C_{T-1}} \right) \\
&+ 2Pd \sqrt{\frac{L}{K(\lambda + L) C_{T-1}}}.
\end{align*}

Thus,

\begin{align*}
d_T &= d_{t' - 1} + 1 \\
&< 1/\alpha + 2P (1 - \alpha) \sqrt{\frac{KL}{\lambda} C_{T-1}} / (\alpha\mu_0) \\
&+ \frac{2Pd}{\alpha\mu_0} \sqrt{\frac{L}{K(\lambda + L) C_{T-1}}}.
\end{align*}

(58)

Combine (55), (57) and (58) we have

\begin{align*}
R(T) &< 2PC_T \sqrt{2dT \log \left( 1 + \frac{TKL}{d\lambda} \right)} \\
&+ \left[ \frac{1}{\alpha} + 2P (1 - \alpha) \sqrt{\frac{KL}{\lambda} C_{T-1}} / (\alpha\mu_0) \right] \lambda d \\
&+ \frac{2Pd}{\alpha\mu_0} \sqrt{\frac{L}{K(\lambda + L) C_{T-1}}} \Delta_{\max}.
\end{align*}

\[ \square \]