# Contextual Conservative Interleaving Bandits

**Kei Takemura** [1]

## Abstract

The performance of a bandit algorithm is usually measured by the cumulative rewards of the actions chosen by the algorithm. However, in many real-world applications, the rewards in each round should be good enough for reasons such as safety and fairness. In this paper, we investigate the contextual conservative interleaving bandit problem, which has a performance constraint that requires the chosen actions to be not much worse than given baseline actions in each round. This work is the first to simultaneously consider the following practical situations: (1) multiple actions are chosen in a round, (2) the feature vectors associated with given actions depend on the round, and (3) the performance constraints in each round that depend only on the actions chosen in that round. We propose a meta-algorithm, Greedy on Confidence Widths (GCW), that satisfies the performance constraints with high probability. GCW uses a standard bandit algorithm and achieves minimax optimal regret up to logarithmic factors if the algorithm used is also minimax optimal. We improve the existing analyses for the $C^2$UCB algorithm and the Thompson sampling to combine with GCW. We show that these algorithms achieve near-optimal regret when the feasible sets of given actions are the bases of a matroid. Our numerical experiments on a real-world dataset demonstrate that GCW with the standard bandit algorithms efficiently improves performance while satisfying the performance constraints.

## 1. Introduction

The stochastic multi-armed bandit (MAB) problem and its extensions have been studied as sequential decision-making problems under uncertainty. In the MAB problem, a learner iterates the following process $T$ times: The learner chooses an action from given actions and then obtains a reward for the chosen action. The learner aims to maximize the sum of rewards. The central challenge of this problem is the exploration-exploitation trade-off due to the lack of prior knowledge about the rewards. The MAB problem has been generalized to many directions to model more realistic situations of real-world applications. For example, the combinatorial semi-bandit (CS) problem (Gai et al., 2012; Kveton et al., 2014; 2015; Wang & Chen, 2018) enables the learner to choose multiple actions at once and observe corresponding rewards, the contextual linear bandit (CLB) problem (Abe & Long, 1999; Auer, 2002; Abbasi-Yadkori et al., 2011; Chu et al., 2011) introduces time-dependent actions associated with feature vectors, and the contextual combinatorial semi-bandit (CCS) problem (Qin et al., 2014; Takemura & Ito, 2019; Takemura et al., 2021) generalizes both the CS and the CLB problems.

In recent years, the conservative bandit problems (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020; Khezeli & Bitar, 2020; Moradipari et al., 2020; Katariya et al., 2019) have been studied, where constraints on the learner's performance are introduced. Specifically, the learner additionally observes baseline actions for given actions and has to satisfy performance constraints that require that the learner's performance is not much worse than the performance by the baseline actions. The performance constraints mitigate a drawback that the standard bandit algorithms perform poorly in early rounds.

Unfortunately, these studies are not yet suitable for many real-world applications. In real-world applications such as recommendations, the following properties summarized in Table 1 should often be satisfied simultaneously: (1) the learner chooses multiple actions at once, (2) the feasible and the baseline actions depend on the round, and (3) the performance constraints guarantee the performance of *each* round. However, none of the existing studies addresses all of these properties. In Section 2, we will compare this work with the existing studies on the conservative bandit problems in detail.

This paper investigates the *contextual conservative interleaving bandit* (CCIB) problem, which satisfies the three properties above. Specifically, the CCIB problem is the CCS problem with stage-wise performance constraints. As

[1]NEC Corporation, Tokyo, Japan. Correspondence to: Kei Takemura <kei_takemura@nec.com>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

*Table 1.* Conservative bandit problems.

| Algorithm | Combinatorial action set | Time-dependent feasible and baseline actions | Stage-wise constraints |
|---|---|---|---|
| Conservative UCB (Wu et al., 2016) | | | |
| CLUCB (Kazerouni et al., 2017) | | ✓ | |
| CLUCB2 (Garcelon et al., 2020) | | ✓ | |
| SEGE (Khezeli & Bitar, 2020) | | | ✓ |
| SCLUCB & SCLTS (Moradipari et al., 2020) | | | ✓ |
| iUCB (Katariya et al., 2019) | ✓ | | ✓ |
| **Algorithm 1 (This work)** | ✓ | ✓ | ✓ |

in previous studies on bandit algorithms, we measure the performance of an algorithm by its regret, which is the difference between the sum of the rewards of the optimal actions and that of the algorithm's actions.

Our main contribution is to propose an algorithm that is minimax optimal up to logarithmic factors. Specifically, the proposed algorithm achieves, with probability at least $1-\delta$, $\widetilde{O}(\min(\sqrt{dkT}, d/\Delta)+dk+\min(k\sqrt{dT/n}, dk/(n\Delta)))$ regret while satisfying the constraints, where $d$ is the dimension of the feature vectors, $k$ is the number of actions to be chosen in a round, $T$ is the number of rounds, $n$ is a parameter that controls how conservative the algorithm is, $\Delta$ is a sub-optimality gap, and $\widetilde{O}(\cdot)$ ignores the logarithmic factors in $d$, $k$, $T$, and $1/\delta$. To show the proposed algorithm is almost optimal, we show matching lower bounds by reducing the CCIB problem to the conservative MAB problem.

The proposed algorithm is a meta-algorithm using an algorithm for the CCS problem. Our algorithm interleaves the baseline actions with the actions that the given algorithm recommends to satisfy the performance constraints. The regret by the proposed algorithm can be represented as the sum of the regret due to choosing the baseline actions and that due to choosing the given algorithm's actions. In terms of these high-level ideas, our algorithm and analysis are similar to the iUCB algorithm (Katariya et al., 2019) and its regret analysis for the conservative interleaving bandit (CIB) problem, respectively.[1] However, we cannot use the analysis for the iUCB algorithm to bound the regret of choosing the baseline actions due to the time-dependent actions.

The time-dependent actions make it difficult to bound the regret suffered by the baseline actions. This regret can be bounded by the widths of the confidence intervals of the reward estimates for the baseline and the given algorithm's actions. Since the feasible sets of actions are fixed in the CIB problem, the iUCB algorithm can choose all of the baseline and the given algorithm's actions over multiple rounds. Thus, we can use a standard analysis for UCB-type algorithms to bound the regret. However, we cannot take this approach for the CCIB problem since a feature vector in a round may never appear in future rounds. Therefore, it is necessary to bound the confidence widths for the actions that are *not* chosen by the proposed algorithm since the proposed algorithm cannot choose all the actions of the baseline and the given algorithm's actions.

To overcome this difficulty, the proposed algorithm preferentially chooses actions with large confidence widths from the actions of the baseline and the given algorithm. This technique allows us to bound the confidence widths for the actions that are not chosen by the proposed algorithm using those for the the proposed algorithm's actions. Consequently, we can use existing analyses to bound the regret due to satisfying the performance constraints. Since the proposed algorithm chooses the actions with large confidence widths, a sophisticated analysis is needed to bound the sum of these confidence widths. To the best of our knowledge, our analysis is the first to bound the regret by the confidence widths for the actions *not* chosen by the algorithm. We believe that our technique is useful for sibling problems.

We consider the C$^2$UCB algorithm (Qin et al., 2014) and the (round-wise) Thompson sampling (Takemura & Ito, 2019) as the algorithm that the proposed algorithm uses. We show that these algorithms achieve near-optimal regret bounds for the CCS problem with the feasible actions characterized by *any* matroid. Moreover, we show the first gap-dependent regret bound for the C$^2$UCB algorithm. Specifically, the C$^2$UCB algorithm and the Thompson sampling achieve $\widetilde{O}(\min(\sqrt{dkT}, d/\Delta)+dk)$ and $\widetilde{O}(d^{3/2}\sqrt{kT}+dk)$ regrets, respectively. The first term of the regret bound of the Thompson sampling has a gap from the optimal bound. However, we can also see this gap in the regret bound of Thompson sampling for the CLB problem (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017). Note that the proposed algorithm is minimax optimal if it uses the C$^2$UCB algorithm.

[1] While the iUCB algorithm is not a meta-algorithm, we consider the iUCB algorithm to be a combination of a meta-algorithm and a UCB-type algorithm for ease of presentation.

We evaluate the proposed algorithm through numerical experiments on a real-world dataset. We conduct our experiments in two cases. One is the case where the feasible and the baseline actions are fixed across the rounds. The other is the case where the given actions depend on the round. In the first case, we compare the proposed algorithm with the iUCB algorithm and its variant that uses the feature vectors to estimate the rewards and the confidence intervals. We also compare our algorithm with the baseline actions of the proposed algorithm and some standard bandit algorithms in both cases. In the first case, the proposed algorithm outperforms other algorithms except for the variant of the iUCB algorithm. The variant of the iUCB algorithm is compatible with the proposed algorithm in terms of regret. However, the variant of the iUCB algorithm is unstable in terms of performance constraints. In the second case, the proposed algorithm behaves similarly to the first. In other words, we observe that our algorithm is robust to the changes in the given actions.

## 2. Related Work

### 2.1. Conservative Bandits

The concept of the conservative bandit problems was proposed by Wu et al. (2016). They studied the MAB problem with constraints that in each round, the learner has to satisfy that the cumulative reward until this round is not much worse than the cumulative reward by the baseline. This work was generalized to the CLB problem (Kazerouni et al., 2017; Garcelon et al., 2020). While these studies consider the constraints on cumulative rewards, the performance constraints should guarantee the performance of *each* round in real-world applications. For instance, let us consider a recommender system. If the system recommends items that do not match the target user's preferences at all due to exploration by the bandit algorithms, this user may never use the system again. Furthermore, it is unfair to recommend unfavorable items to one user due to the exploration but favorable items to another.

Khezeli & Bitar (2020) and Moradipari et al. (2020) introduced stage-wise constraints that consider the rewards of each round to the linear bandit problem with a time-independent convex action set. While the stage-wise constraints solve the performance issue in early rounds, these studies strongly rely on the assumption that the action set is time-independent and convex. In other words, we cannot apply these studies to many real-world applications (e.g., the recommender system considered above).

Katariya et al. (2019) studied the CIB problem, which is the CS problem with baseline actions and stage-wise performance constraints. They simultaneously handled the non-convex action set and the stage-wise constraints by introducing the problem in which the learner chooses multiple actions in a round. Moreover, the learner must often choose multiple actions simultaneously in real-world applications. Note that the CIB and CCIB problems denote the problems of maximizing the sum of rewards, while the CS and CCS problems cover some classes of non-linear functions of the rewards.

While the CIB problem assumes that the feasible and the baseline actions do not depend on the round, the feasible and the baseline actions often depend on the round in real-world applications. In a news article recommendation, for example, news articles would change over the rounds as old news articles are deleted or the latest news articles are added (Li et al., 2010; Wang et al., 2017). The given actions correspond to the news articles in this example. Note that the feasible and the baseline actions could depend on the round, even if the given actions are fixed over the rounds.

### 2.2. Contextual Combinatorial Semi-Bandits

Qin et al. (2014) proposed the CCS problem and the C$^2$UCB algorithm. Takemura & Ito (2019) and Takemura et al. (2021) improved the regret bound by the C$^2$UCB algorithm. In particular, Takemura et al. (2021) showed that the C$^2$UCB algorithm is minimax optimal up to logarithmic factors when the partition matroid characterizes the feasible actions. Our analysis of the C$^2$UCB algorithm is an extension of the analysis by Takemura et al. (2021).

Takemura & Ito (2019) proposed Thompson sampling algorithms for the CCS problem was proposed and showed that this algorithm achieves $\widetilde{O}(\max(\sqrt{d}, \sqrt{k})d\sqrt{kT})$ regret. Unlike the C$^2$UCB algorithm, the Thompson sampling algorithms in the CCS problem with the feasible actions characterized by matroids have not been studied. In addition, a gap-dependent bound of the Thompson sampling algorithms for the CCS problem is not known.

## 3. Problem Setting

### 3.1. Contextual Conservative Interleaving Bandits

We formally define the CCIB problem. This problem consists of $T$ rounds, and a learner chooses a set of $k$ actions from a given family of action sets in each round. Each action, called *arm*, is associated with a $d$-dimensional feature vector. Let $N$ denote the number of given arms, and each arm is indexed by an integer in $[N] := \{1, 2, \ldots, N\}$. Note that the learner knows the above parameters in advance.

The learner progresses through each round as follows. At the beginning of the $t$-th round, a learner observes the feature vectors $\{x_t(i)\}_{i \in [N]} \subseteq \mathbb{R}^d$ of the arms. In addition, the learner observes a family $\mathcal{S}_t \subseteq 2^{[N]}$ of feasible sets of arms and a feasible set $B_t \in \mathcal{S}_t$ called baseline arms. Then, the

learner chooses a feasible set $I_t \in \mathcal{S}_t$. We assume that each feasible set of arms is of size $k$, i.e., $\forall I \in \mathcal{S}_t$, $|I| = k$. At the end of the round, the learner obtains the rewards $\{r_t(i)\}_{i \in I_t}$, where for all $i \in I_t$, $r_t(i) = \theta^\top x_t(i) + \eta_t(i)$ for some unknown parameter $\theta \in \mathbb{R}^d$ and a random noise $\eta_t(i) \in \mathbb{R}$. We will introduce the assumption on $\eta_t(i)$ in Section 3.2.

We evaluate the performance of the learner by the regret $R(T)$, which is defined as

$$R(T) = \sum_{t \in [T]} \left( \sum_{i \in I_t^*} \mu_t(i) - \sum_{i \in I_t} \mu_t(i) \right),$$

where $\mu_t(i) = \theta^\top x_t(i)$ for all $i \in [N]$ and $t \in [T]$, and $I_t^* \in \operatorname{argmax}_{I \in \mathcal{S}_t} \sum_{i \in I} \mu_t(i)$ for all $t \in [T]$. The learner aims to minimize the regret, which is equivalent to maximizing $\sum_{t \in [T]} \sum_{i \in I_t} \mu_t(i)$.

Compared to the standard bandit problems, the CCIB problem additionally introduces stage-wise performance constraints as in the CIB problem (Katariya et al., 2019). These constraints require that chosen arms' performance in each round should not be much worse than the baseline arms' performance of the round. More precisely, with high probability, there exists a bijection $\rho_t \colon I_t \to B_t$ such that

$$\sum_{i \in I_t} \mathbb{1}(\mu_t(i) < \mu_t(\rho_t(i))) \leq m \tag{1}$$

for all $t \in [T]$, where $m$ is a non-negative parameter that represents how $I_t$ should be conservative.[2] Our constraint allows aggressive explorations when $m$ is large. Note that choosing $B_t$ satisfies this constraint.

Several existing studies (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020; Moradipari et al., 2020) define a performance constraint using the sum of obtained rewards. An adaptation of this constraint to our problem is that

$$\sum_{i \in I_t} \mu_t(i) \geq (1 - \alpha) \sum_{i \in B_t} \mu_t(i) \tag{2}$$

for a fixed $\alpha \in [0, 1]$ for all $t \in [T]$, with high probability. In general, this constraint is not directly comparable to our constraint. We show this fact in Appendix B.1. However, we show that constraint (1) implies constraint (2) under a reasonable assumption, which will be discussed in Section 3.2.

### 3.2. Assumptions

We introduce several assumptions on the CCIB problem and define a few parameters of the problem.

[2]We assume that the parameters $k$ and $m$ are time-independent for simplicity. However, our algorithm and analysis can be easily extended to the time-dependent parameters.

First, we discuss the assumptions of the feature vectors, the unknown parameter, the rewards, and their noises. Recall that $\mu_t(i) = \theta^\top x_t(i)$ and $r_t(i) = \mu_t(i) + \eta_t(i)$. We assume the following standard assumptions in the CCS problem (Qin et al., 2014; Takemura & Ito, 2019; Takemura et al., 2021) and the conservative bandit problems (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020):

**Assumption 3.1.** The feature vectors $\{x_t(i)\}_{i \in [N]}$ are determined by oblivious adversary.

**Assumption 3.2.** The learner knows $\|\theta\|_2 \leq M$ and $\|x_t(i)\|_2 \leq L$ for all $i \in [N]$ and $t \in [T]$.

**Assumption 3.3.** The mean reward $\mu_t(i)$ satisfies $\mu_t(i) \in [0, 1]$ for all $i \in [N]$ and $t \in [T]$.

**Assumption 3.4.** The noise sequence $\{\eta_t(i)\}_{i \in I_t, t \in [T]}$ is conditionally $R$-sub-Gaussian, i.e,

$$\mathbb{E}\left[\exp(\lambda \eta_t(i)) \mid \{x_s(j)\}_{j \in I_s, s \in [t]}, \{\eta_s(j)\}_{j \in I_s, s \in [t-1]}\right] \leq \exp(\lambda^2 R^2 / 2)$$

for all $\lambda \in \mathbb{R}$, $i \in I_t$ and $t \in [T]$.

Note that Assumption 3.4 implies that the mean of $\eta_t(i)$ is zero for all $i \in I_t$ and $t \in [T]$.

Next, we define the requirements for the algorithm used by the proposed meta-algorithm.

**Assumption 3.5.** The algorithm used by the proposed meta-algorithm works for the CCS problem (i.e., the CCIB problem without the performance constraints). The algorithm has an internal model for the estimation of rewards. In a round $t$, the algorithm runs as follows:

1. It estimates the rewards of given arms based on observed feature vectors.
2. It chooses arms by solving $\operatorname{argmax}_{I \in \mathcal{S}_t} r'_t(i)$, where $\{r'_t(i)\}_{i \in [N]}$ is estimation of the rewards.
3. It plays the chosen arms and updates the internal model using the feature vectors, the chosen arms, and the observed rewards.

Furthermore, the algorithm can retry the second step with another set $\mathcal{S}'_t$ of feasible actions as long as it is before executing the third step.

Note that the standard bandit algorithms such as the UCB-type algorithms, the Thompson sampling, and the ($\varepsilon$-)greedy algorithm satisfy Assumption 3.5.

Next, we discuss the family of feasible sets of arms, i.e., $\mathcal{S}_t$. As in the CIB problem (Katariya et al., 2019), we focus on exchangeable set:

**Definition 3.6** (Exchangeable set)**.** Given a set $E$, a family $\mathcal{S} \subseteq 2^E$ is exchangeable if for any two sets $I_1, I_2 \in \mathcal{S}$, there exists a bijection $\rho : I_1 \to I_2$ such that

$$\forall J \subseteq I_1,\ (I_1 \setminus J) \cup \{\rho(i) \mid i \in J\} \in \mathcal{S}. \tag{3}$$

**Assumption 3.7.** The family of feasible sets of arms, $\mathcal{S}_t$, is exchangeable for all $t \in [T]$.

**Assumption 3.8.** For any $t \in [T]$ and $I_1, I_2 \in \mathcal{S}_t$, the learner knows a bijiection $\rho : I_1 \to I_2$ that satisfies (3).

The set $E$ in Definition 3.6 corresponds to $[N]$ in our problem. Note that if $\mathcal{S}_t$ is the bases of a uniform matroid or a partition matroid, it is known how to construct the bijection that satisfies (3) (Katariya et al., 2019).

We discuss how large the class of exchangeable sets is. Katariya et al. (2019) pointed out that the set of the bases of a strongly base-orderable matroid is exchangeable, but the converse remains an open question. This paper solves this question in the affirmative, i.e., we show that every exchangeable set is the set of the bases of some strongly base-orderable matroid. We defer our proof to Appendix C. It is known that the strongly base-orderable matroid includes the uniform matroid, the partition matroid, and the transversal matroid.

Next, we discuss the lower bound of the parameter $m$. Intuitively, small $m$ makes the CCIB problem difficult. In fact, when $m = 0$, no algorithm can achieve sub-linear regret while satisfying our performance constraint. We defer our proof to Appendix E.1. To obtain sub-linear regret, we assume the following:

**Assumption 3.9.** The parameter $m$ satisfies $m \geq 1$.

Finally, we introduce a reasonable assumption for a reduction from the constraint (2) to constraint (1). We defer our proof to Appendix B.2.

**Assumption 3.10.** Let $r_\ell = \min_{t\in[T]} \min_{i\in B_t} \mu_t(i)$. Then, the parameters $k$, $m$, $\alpha$ and $r_\ell$ satisfy $\alpha r_\ell \geq m/(k-m)$.

Note that we do not need Assumption 3.10 to satisfy constraint (1). Since we have a reduction, we will focus on constraint (1) in the rest of this paper.

## 4. Proposed Algorithm

In this section, we propose an algorithm for the CCIB problem. Our algorithm, described in Algorithm 1, is the first to deal with time-dependent feature vectors and stage-wise performance constraints, as discussed in Section 1. Since our algorithm is a meta-algorithm, it requires a base algorithm $\mathcal{A}$ for the CCS problem as input.

At the beginning of the $t$-th round, our algorithm observes the feature vectors $\{x_t(i)\}_{i\in[N]}$, feasible sets $\mathcal{S}_t$, and baseline arms $B_t$ (line 3). Then, our algorithm receives the arms

**Algorithm 1** Greedy on confidence widths

**Input:** $\lambda > 0$, $\delta \in (0,1)$, $n \in \mathbb{N}$, and an algorithm $\mathcal{A}$.
1: $V_0 \leftarrow \lambda I$ and $b_0 \leftarrow \mathbf{0}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Observe $\{x_t(i)\}_{i\in[N]}$, $\mathcal{S}_t$, and $B_t$.
4: $\hat{I}_t \in \operatorname{argmax}_{I\in\mathcal{S}_t} \sum_{i\in I} r'_t(i)$, where $\{r'_t(i)\}_{i\in[N]}$ is the estimation of rewards by $\mathcal{A}$.
5: Define $\bar{r}_t(i)$ for all $i \in [N]$ according to (4).
6: $\bar{I}_t \in \operatorname{argmax}_{I\in\mathcal{S}_t} \sum_{i\in I} \bar{r}_t(i)$.
7: Let $\bar{\rho}_t : \bar{I}_t \to \hat{I}_t$ be the bijection that satisfies (3).
8: $\bar{I}_t^0 \leftarrow \bar{I}_t$ and $J_t^0 \leftarrow \emptyset$.
9: **for** $\ell = 1, 2, \ldots, n$ **do**
10: $i_t^\ell \in \operatorname{argmax}_{i\in\bar{I}_t^{\ell-1}} \max(c_t(i), c_t(\bar{\rho}_t(i)))$.
11: $\bar{I}_t^\ell \leftarrow \bar{I}_t^{\ell-1} \setminus \{i_t^\ell\}$.
12: $j_t^\ell \leftarrow \operatorname{argmax}_{j\in\{i_t^\ell, \bar{\rho}_t(i_t^\ell)\}} c_t(j)$.
13: $J_t^\ell \leftarrow J_t^{\ell-1} \cup \{j_t^\ell\}$.
14: **end for**
15: Choose $I_t \in \mathcal{S}_t$ such that $|I_t \setminus \bar{I}_t| \leq n$, $J_t^n \subseteq I_t \subseteq \hat{I}_t \cup \bar{I}_t$, and $i \in I_t$ if and only if $\bar{\rho}_t(i) \notin I_t$ for all $i \in \bar{I}_t \setminus \hat{I}_t$.
16: Play $I_t \in \mathcal{S}_t$ and observe the rewards $\{r_t(i)\}_{i\in I_t}$.
17: Feed $\hat{I}_t \cap I_t$ and the corresponding rewards to $\mathcal{A}$, and update $\mathcal{A}$ as if it has chosen $\hat{I}_t \cap I_t$.
18: $V_t \leftarrow V_{t-1} + \sum_{i\in I_t} x_t(i) x_t(i)^\top$.
19: $b_t \leftarrow b_{t-1} + \sum_{i\in I_t} r_t(i) x_t(i)$.
20: **end for**

$\hat{I}_t$ that the base algorithm $\mathcal{A}$ recommends (line 4). In the following and our analysis, we call $\hat{I}_t$ *recommended choice*.

Then, the proposed algorithm starts to interleave the recommended arms with the baseline arms. At the first step of our interleaving, the proposed algorithm obtains safe arms $\bar{I}_t$ such that there exists a bijection $\rho_t : \bar{I}_t \to B_t$ that satisfies $\mu_t(i) \geq \mu_t(\rho_t(i))$ for all $i \in \bar{I}_t$ with high probability. To obtain the safe arms, we define the following rewards and solve the maximization problem (lines 5 and 6):

$$\bar{r}_t(i) = \begin{cases} \hat{r}_t(i) & \text{if } i \in B_t \\ \check{r}_t(i) & \text{if } i \in \hat{I}_t \setminus B_t \\ -\infty & \text{otherwise} \end{cases}, \tag{4}$$

where $\hat{r}_t(i) = \hat{\theta}_t^\top x_t(i) + c_t(i)$, $\check{r}_t(i) = \hat{\theta}_t^\top x_t(i) - c_t(i)$, $\hat{\theta}_t = V_{t-1}^{-1} b_{t-1}$, $c_t(i) = \beta_t(\delta)\sqrt{x_t(i)^\top V_{t-1}^{-1} x_t(i)}$, and $\beta_t(\delta) = \widetilde{O}(\sqrt{\min(\log(N), d)} + M\sqrt{\lambda})$ for all $i \in [N]$. The formal definition of $\beta_t(\delta)$ is described in Appendix D.1. Note that $\bar{I}_t \subseteq \hat{I}_t \cup B_t$ because $B_t$ is a feasible solution with a finite objective value. Note also that since $\mathcal{S}_t$ is the bases of a matroid, this optimization problem can be solved efficiently by the greedy algorithm (Korte & Vygen, 2018).

Then, the proposed algorithm chooses the arms from $\hat{I}_t \cup \bar{I}_t$ for the exploration. Our algorithm needs a positive integer

$n$, which controls the maximum number of arms for the exploration. The algorithm preferentially chooses $n$ arms with large confidence widths as $J_t^n$ (lines 8–14). Then, the algorithm chooses the playing arms $I_t$ that includes $J_t^n$ (line 15). Note that one can choose any set of arms that satisfies all conditions in line 17 of Algorithm 1. For example, the set $(\bar{I}_t \setminus \{j \in J_t^n \cap (\hat{I}_t \setminus \bar{I}_t) \mid \bar{\rho}_t^{-1}(j)\}) \cup (J_t^n \cap (\hat{I}_t \setminus \bar{I}_t))$ satisfies the conditions. If $|J_t^n \cap (\hat{I}_t \setminus \bar{I}_t)| < n$ and $\hat{I}_t \setminus (J_t^n \cup \bar{I}_t) \neq \emptyset$, we can further exchange the arms in $\hat{I}_t \setminus (J_t^n \cup \bar{I}_t)$ for the corresponding baseline arms.

At the end of the $t$-th round, the proposed algorithm and the base algorithm $\mathcal{A}$ update their reward estimation by the observed rewards (lines 17–19). This update of $\mathcal{A}$ can be seen as a retry of choosing arms by $\mathcal{A}$. In fact, it holds that $\hat{I}_t \cap I_t \in \operatorname{argmax}_{I \in \mathcal{S}_t'} \sum_{i \in [N]} r_t'(i)$ for some $\mathcal{S}_t'$ (see our proof of Lemma 5.2 for details). Therefore, the base algorithm $\mathcal{A}$ runs on an instance of the CCS problem. In other words, we can use existing regret bounds of $\mathcal{A}$ to bound the regret by $\hat{I}_t \cap I_t$.

To bound the regret by the proposed algorithm, the central part of our analysis (and the analysis by Katariya et al. (2019)) is to bound the regret by $\{I_t \setminus \hat{I}_t\}_{t \in [T]}$, i.e., bounding the penalty by satisfying our performance constraint. We call the set $I_t \setminus \hat{I}_t$ *conservative choice*. Note that $I_t \setminus \hat{I}_t \subseteq B_t$. The regret by the conservative choices can be bounded by the confidence widths for the conservative choices and those for the recommended choices.

For the CIB problem (i.e., $\{x_i(i)\}_{i \in [N]}$, $\mathcal{S}_t$, and $B_t$ are fixed over the round)[3], Katariya et al. (2019) showed that the confidence widths by the iUCB algorithm for the conservative choices can be bounded by those for the recommended choices. Since $\mathcal{S}_t$ fixed, the iUCB algorithm chooses the conservative and the recommended arms using multiple rounds and it is possible to use a standard analysis for UCB-type algorithms to bound the regret. However, we cannot use this approach for the CCIB problem because of the contextual setting. First, we cannot bound the confidence width for the conservative choices by that for the recommended choices since the baseline arms may change. Second, our algorithm may not be able to choose the recommended choices in other rounds because these feature vectors may not appear in other rounds.

To solve these problems, our algorithm chooses $J_t^n$. As discussed in Section 1, choosing $J_t^n$ enables us to bound the confidence widths for the arms that are not chosen by our algorithm using those for the playing arms $I_t$. Then, we can utilize the existing analyses for the standard bandit algorithms. We formally show this fact in the next section.

[3]Though Katariya et al. (2019) dose not consider the case that the feature vectors associated with the arms, the iUCB algorithm can be applied to the problem with time-independent feature vectors by ignoring the feature vectors.

## 5. Regret Analysis

In this section, we introduce and prove our regret bound of the proposed algorithm. Let $\mathcal{S}_t^* = \operatorname{argmax}_{I \in \mathcal{S}_t} \sum_{i \in I} \mu_t(i)$. Throughout our analysis, we fix $I_t^* \in \mathcal{S}_t^*$ arbitrarily for all $t \in [T]$ and assume that $\forall t \in [T], \mathcal{S}_t \neq \mathcal{S}_t^*$ for ease of presentation. Note that we have no regret when $\mathcal{S}_t = \mathcal{S}_t^*$. Then, we define the following sub-optimality gaps:[4] $\Delta_t(i,j) = \mu_t(j) - \mu_t(i)$ and $\Delta = \min_{t \in [T], I \in \mathcal{S}_t \setminus \mathcal{S}_t^*, i \in I, i^* \in I_t^* : \Delta_t(i,i^*) > 0} \Delta_t(i, i^*)$.

Using our gap $\Delta$, we introduce our regret bound:

**Theorem 5.1.** *Let $\kappa = \min(\log(N), d)$, $\nu = \log(kT)$ and $\iota = \min(\log(NkT/\delta), d\log(kT/\delta))$. If $\lambda = (R/M)^2\kappa$ and $n \leq m$, with probability at least $1-\delta$, Algorithm 1 satisfies performance constraint (1) for all $t \in [T]$ and achieves the following regret bound:*

$$R(T) = R_{\mathcal{A}}(T) + O\left(\min\left(\sqrt{dkT\nu\iota}, d\nu\iota/\Delta\right) + dk\nu + \min\left(k\sqrt{dT\nu\iota/n}, dk\nu\iota/(n\Delta)\right)\right),$$

*where $R_{\mathcal{A}}(T)$ is the regret bound of the base algorithm $\mathcal{A}$ for the CCS problem.*

Our regret bound consists of the regret bound of the base algorithm $\mathcal{A}$ and the penalty for satisfying the performance constraints. We will show in Section 6 that the penalty term is minimax optimal up to logarithmic factors. Furthermore, the penalty term matches the regret bound of the iUCB algorithm by Katariya et al. (2019) when we apply the proposed algorithm to the CIB problem.

In the remainder of this section, we prove Theorem 5.1 and the regret bounds for the two algorithms used as the base algorithm $\mathcal{A}$. We defer all missing proofs to Appendix D. Recall that $\hat{I}_t$ is the recommended choice (line 4 of Algorithm 1), and $I_t \setminus \hat{I}_t$ is the conservative choice. Let $\hat{\rho}_t^* : \hat{I}_t \to I_t^*$ be the bijection satisfying (3). Let $\rho_t^* : I_t \to I_t^*$ denote $\rho_t^*(i) = \hat{\rho}_t^*(i)$ if $i \in \hat{I}_t$ and $\rho_t^*(i) = \hat{\rho}_t^*(\bar{\rho}_t(i))$ otherwise. Let $R_t(i) = \mu_t(\rho_t^*(i)) - \mu_t(i)$ for $i \in I_t \setminus \hat{I}_t$.

Our analysis decomposes the regret as follows and bounds them separately:

$$R(T) = \sum_{t \in [T]} \left( \sum_{i \in I_t \cap \hat{I}_t} R_t(i) + \sum_{i \in I_t \setminus \hat{I}_t} R_t(i) \right). \quad (5)$$

The former is the regret of the recommended choices, and the latter is the regret of the conservative choices. We have the following lemma for the regret by the recommended choices:

[4]If the feature vectors are fixed over the rounds, we can define alternative sub-optimality gaps $\Delta_{e,\min}$ and $\Delta_{e^*,\min}^*$ by Katariya et al. (2019). In this case, we have $\Delta = \min(\Delta_{e,\min}, \Delta_{e^*,\min}^*)$.

**Lemma 5.2.** *We have* $\sum_{t\in[T]}\sum_{i\in I_t\cap\hat{I}_t} R_t(i) \le R_{\mathcal{A}}(T)$.

As a common tool for our analysis, we bound the estimation errors of the rewards as in previous work (Qin et al., 2014; Takemura & Ito, 2019; Takemura et al., 2021). Let $\|x\|_A$ denote $\sqrt{x^\top Ax}$ for all $x \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ such that $A$ is positive definite. Then, we have the following:

**Lemma 5.3.** *Suppose that* $\delta \in (0,1)$. *We have, with probability at least* $1-\delta$, *for all* $t \in [T]$ *and* $i \in [N]$,

$$|\mu_t(i) - \hat{\theta}_t^\top x_t(i)| \le \beta_t(\delta)\|x_t(i)\|_{V_{t-1}^{-1}}. \tag{6}$$

In our proof of Theorem 5.1, we assume that (6) holds.

### 5.1. Regret by Conservative Choices

Similar to the analysis by Katariya et al. (2019), we can bound the regret of an arm in the conservative choices by the sum of the confidence widths for that arm and the corresponding arm in the recommended choices:

**Lemma 5.4.** *For any* $t \in [T]$ *and any* $i \in I_t \setminus \hat{I}_t$, *we have* $\mu_t(\rho_t^*(i)) - \mu_t(i) \le 2(c_t(i) + c_t(\bar{\rho}_t(i)))$.

Since the confidence widths of $J_t^n$ is larger than those of $(\bar{I}_t \cup \hat{I}_t) \setminus J_t^n$, we can bound the regret by the sum of the confidence widths of $J_t^n$. However, to bound such large confidence widths, we need two additional theoretical tools.

First, we show that the number of times the confidence widths for $J_t^n$ are greater than the sub-optimality gap can be bounded. Recall that $j_t^\ell$ is defined in line 15 of Algorithm 1 for all $\ell \in [n]$.

**Lemma 5.5.** *Let* $\gamma_\ell = \sqrt{\ell}\Delta/(4\beta_T(\delta))$ *for all* $\ell \in [n]$. *Then, for any* $\ell \in [n]$, *we have* $\sum_{t\in[T]} \mathbb{1}(c_t(j_t^\ell) \ge \Delta/4) \le 2d\log(1 + L^2kT/(d\lambda))/\min(\gamma_\ell^2, 1)$.

Second, we show that the conservative choice suffers no regret if their confidence widths are smaller than the sub-optimality gap. Recall that $\hat{\rho}_t^* : \hat{I}_t \to I_t^*$ and $\bar{\rho}_t(i) : \bar{I}_t \to \hat{I}_t$ are the bijections satisfying (3).

**Lemma 5.6.** *For any* $i \in \bar{I}_t$, *if* $c_t(i) + c_t(\bar{\rho}_t(i)) < \Delta/2$, *we have* $\mu_t(i) = \mu_t(\bar{\rho}_t(i)) = \mu_t(\hat{\rho}_t^*(\bar{\rho}_t(i)))$.

We are ready to bound the regret due to the conservative choices. The regret can be rewritten as

$$\sum_{t\in[T]}\sum_{i\in I_t\setminus\hat{I}_t} R_t(i) = \sum_{t\in[T]}\sum_{i\in I_t\setminus(\hat{I}_t\cup J_t^n)} R_t(i) \tag{7}$$

$$+ \sum_{t\in[T]}\sum_{i\in J_t^n\setminus\hat{I}_t} R_t(i). \tag{8}$$

First, we bound the regret due to the right-hand side of (7). Let $\bar{I}_t' = I_t \setminus (\hat{I}_t \cup J_t^n)$. Since $\beta_T(\delta) \ge \beta_t(\delta)$ for all $t \in [T]$ and $c_t(i) + c_t(\bar{\rho}_t(i)) \le 2c_t(j_t^n)$ for all $I_t \setminus J_t^n$ and $t \in [T]$, by Lemma 5.4 and Lemma 5.6, we obtain

$$\sum_{t\in[T]}\sum_{i\in\bar{I}_t'} R_t(i) \le \sum_{t\in\Phi}\sum_{i\in\bar{I}_t'} R_t(i) + \sum_{t\in\Psi}\sum_{i\in\bar{I}_t'} R_t(i)$$
$$\le 2k\beta_T(\delta)\sum_{t\in\Phi}\|x_t(j_t^n)\|_{V_t^{-1}} + k|\Psi|,$$

where $\Phi = \{t \in [T] \mid c_t(j_t^n) \ge \frac{\Delta}{4}, \|x_t(j_t^n)\|^2_{\widetilde{V}^{-1}_{t-1,n}} \le \frac{1}{n}\}$, $\Psi = \{t \in [T] \mid \|x_t(j_t^n)\|^2_{\widetilde{V}^{-1}_{t-1,n}} > \frac{1}{n}\}$, and $\widetilde{V}_{t,\ell} = \lambda I + \sum_{s\in[t]}\sum_{j\in J_t^\ell} x_s(j)x_s(j)^\top$ for all $\ell \in [n]$ and $t \in [T]$. We consider the rounds in $\Phi$. Let $\xi = \log(1 + L^2kT/(d\lambda))$. Then, we have

$$\sum_{t\in\Phi}\|x_t(j_t^n)\|_{V_{t-1}^{-1}} \le \frac{1}{n}\sum_{t\in\Phi}\sum_{j\in J_t^n}\min\left(\frac{1}{\sqrt{n}}, \|x_t(j)\|_{\widetilde{V}^{-1}_{t-1,n}}\right)$$
$$\le \frac{1}{n}\sqrt{n|\Phi|\sum_{t\in\Phi}\sum_{j\in J_t^n}\min\left(\frac{1}{n}, \|x_t(j)\|^2_{\widetilde{V}^{-1}_{t-1,n}}\right)}$$
$$\le \min\left(\sqrt{\frac{2dT\xi}{n}}, \frac{2d}{\sqrt{n}\min(\gamma_n, 1)}\xi\right),$$

where the first inequality is derived from the facts that $V_t \succeq \widetilde{V}_{t,n}$ and $c_t(j_t^n) \le c_t(j_t^\ell)$ for all $\ell \in [n]$ and $t \in [T]$, the second inequality holds by the Cauchy-Schwarz inequality, and the last inequality is obtained by Lemma A.3 (Lemma 5 by Takemura et al. (2021)) and Lemma 5.5.

For the rounds in $\Psi$, by a similar discussion above, we have $k|\Psi| \le \frac{k}{n}\sum_{t\in[T]}\sum_{j\in J_t^n}\mathbb{1}(\|x_t(j)\|^2_{\widetilde{V}^{-1}_{t-1,n}} > \frac{1}{n}) \le 2dk\xi$, where the last inequality follows from Lemma A.3.

Next, we bound the regret due to (8). Using Lemma 5.4, we have $R_t(j) \le 2\beta_t(\delta)\|x_t(j)\|_{V_{t-1}^{-1}}$ for all $j \in J_t^n$ and $t \in [T]$. Thus, by a similar analysis for the C$^2$UCB algorithm, we have the following bound (see Appendix D.7 for details): $\sum_{t\in[T]}\sum_{i\in J_t^n\setminus\hat{I}_t} R_t(i) = \widetilde{O}(\min(\sqrt{dnT\nu\iota}, \frac{d\nu\iota}{\Delta}) + dn\nu)$.

Finally, combining the above discussions, we obtain the desired result.

### 5.2. Performance Guarantee

The rest of our proof of Theorem 5.1 is to show that the set $\{I_t\}_{t\in[T]}$ of the playing arms satisfies our performance constraint (1) with high probability. To show this performance guarantee, we bound the number of arms to be chosen for the exploration. To prove the desired result, we use the following lemma:

**Lemma 5.7.** *Let* $E$ *be the ground set of a matroid,* $\mathcal{A}$ *be the set of the bases of the matroid, and* $A \in \mathcal{A}$ *be a basis. Let* $r : E \to \mathbb{R}$ *be a reward function and* $A^*$ *be a basis such that* $A^* \in \operatorname{argmax}_{I\in\mathcal{A}}\sum_{i\in I} r(i)$. *Then, there exists*

*a bijection $\rho : A^* \to A$ such that $r(i) \geq r(\rho(i))$ for all $i \in A^*$. In addition, we have $\rho(i) = i$ for all $i \in A^* \cap A$.*

We fix $t \in [T]$ arbitrarily. Let $\rho'_t : \bar{I}_t \to B_t$ be a bijection by Lemma 5.7 with $\{\bar{r}_t(i)\}_{i\in[N]}$. Then, since (6) holds, for any $i \in \bar{I}_t$, we have $\mu_t(i) \geq \bar{r}_t(i) \geq \bar{r}_t(\rho'_t(i)) \geq \mu_t(\rho'_t(i))$. Hence, it is sufficient to compare $I_t$ with $\bar{I}_t$. By the construction of $I_t$, we always have $|I_t \setminus \bar{I}_t| \leq n$. This fact and $n \leq m$ imply that $I_t$ satisfies the constraint.

### 5.3. Near-Optimal Regret Bounds for the CCS Problem

In this subsection, we briefly introduce our analysis for the C$^2$UCB algorithm and the Thompson sampling. A full proof is provided in Appendix G and Appendix H. Note that we consider the CCS problem and the definition of regret is the same as the regret of the CCIB problem.

The key component of our proof is the bijection $\rho^*_t : I_t \to I^*_t$ by Lemma 5.7 with the reward estimates. These estimates $\{r'_t(i)\}_{i\in[N]}$ correspond to the upper confidence bounds and the rewards calculated by the sampled parameter in the C$^2$UCB algorithm and the Thompson sampling, respectively. Recall that $R_t(i) = \mu_t(\rho^*(i)) - \mu_t(i)$. Then, we can decompose the regret as follows: $R(T) = \sum_{t\in[T]} \left(\sum_{i\in I_t\setminus J_t} R_t(i) + \sum_{i\in J_t} R_t(i)\right)$, where $J_t = \{i \in I_t \mid \|x_t(i)\|_{V_{t-1}^{-1}} \geq 1/\sqrt{k}\}$. The former term can be bounded by a variant of the existing analyses because we have $r'_t(i) \geq r'_t(\rho^*_t(i))$ for all $i \in I_t$. The latter term can be bounded by the fact that $\sum_{t\in[T]} |J_t| = \widetilde{O}(dk)$.

## 6. Lower Bounds

In this section, we show the following theorem:

**Theorem 6.1.** *Suppose that there exist an algorithm and some $m \leq k/(4e+2)$ such that the algorithm satisfies the constraint (1) in expectation for any environment. Then, for any sub-optimality gap $\Delta \in \left[\sqrt{\frac{d-1}{4mT}}, \sqrt{\frac{d-1}{4m}}\right]$, there is some environment such that $\mathbb{E}[R_T] > \frac{(d-1)k}{(32e+16)m\Delta}$.*

We defer the omitted proofs in this section to Appendix E.2. Since substituting $\Delta = \Theta(\sqrt{d/(mT)})$ leads to a gap-independent bound, we focus on the gap-dependent bound.

To prove Theorem 6.1, we consider instances of the CCIB problem constructed by these of the conservative MAB (CMAB) problem. More precisely, we consider $k$ rounds of the CMAB problem as a round of the CCIB problem. Using the standard basis $\{e_i\}_{i\in[d]}$ as the feature vectors $\{x_t(i)\}_{i\in[d]}$, the CCIB problem reduces to the $d$-armed CMAB problem. To represent $k$ rounds of the CMAB problem as a round of the CCIB problem, it is sufficient to prepare $dk$ arms as $x_t(i + dj) = e_i$ for $i \in [d]$ and $j \in [k]$.

We show that an algorithm for the CCIB problem can be applied to the CMAB problem. The algorithm knows the baseline arms of all rounds in advance because the baseline arm of the CMAB problem is fixed over the rounds. The algorithm can choose arbitrary $k$ arms in a round. Then, one can play the chosen arms in any order in the CMAB problem. Finally, the algorithm observes $k$ rewards and proceeds to the next round.

The remainder of our proof of Theorem 6.1 is to obtain a lower bound of the CMAB problem. Using a variant of the technique by Wu et al. (2016) for the lower bound, we obtain the following bound:

**Theorem 6.2.** *Let $R^{\mathrm{CMAB}}_{T'}$ denote the regret of the $N$-armed CMAB problem with $T'$ rounds. Let $i^{\mathrm{base}}$ be the baseline arm and $i_t$ be the arm chosen in round $t$ by an algorithm. Assume that some algorithm and some positive integer $k$ satisfy that $\mathbb{E}\left[\sum_{t'\in[kt]} \mathbb{1}\left(\bar{r}(i^{\mathrm{base}}) > \bar{r}(i_{t'})\right)\right] \leq \alpha k t$ for all positive integer $t$ such that $kt \leq T'$ for any environment. Then, for any sub-optimality gap $\Delta \in \left[\sqrt{\frac{N-1}{4\alpha T'}}, \sqrt{\frac{N-1}{4\alpha k}}\right]$, there is some environment such that $\mathbb{E}\left[R^{\mathrm{CMAB}}_{T'}\right] > \frac{N-1}{(32e+16)\alpha\Delta}$.*

Substituting $T' = kT$ and $\alpha = m/k$ in Theorem 6.2, we finish the proof of Theorem 6.1.

We discuss our lower bounds. In contrast to our high-probability upper bound, our lower bounds assume that the algorithm satisfies the performance constraints in expectation. However, these lower bounds cover the proposed algorithm. In fact, the proposed algorithm satisfies the performance constraints in expectation if $n = m/2$ and $\delta = \min(n/k, 1/T)$. Note that the expected regret, in this case, has the same regret bound in Theorem 5.1 up to logarithmic factors. Therefore, our lower bounds imply that the proposed algorithm is almost optimal.

## 7. Numerical Experiments

In this section, we experimentally evaluate the performance of the proposed algorithm compared to some existing algorithms using the Book-Crossing dataset (Ziegler et al., 2005). We describe the details of the experimental setup in Appendix F. We conducted our experiments in two settings. First, we considered the case where the feature vectors, the feasible sets of given arms, and the baseline arms are fixed over the rounds. Second, we considered the case where these three components change two times.

Our experiments used the proposed algorithm combined with the C$^2$UCB algorithm (Qin et al., 2014) (GCW + C$^2$UCB) and that combined with the Thompson sampling (Takemura & Ito, 2019) (GCW + TS). We compared the proposed algorithm with the weighted regularized matrix factorization (Hu et al., 2008) as the baseline, the $\varepsilon$-greedy

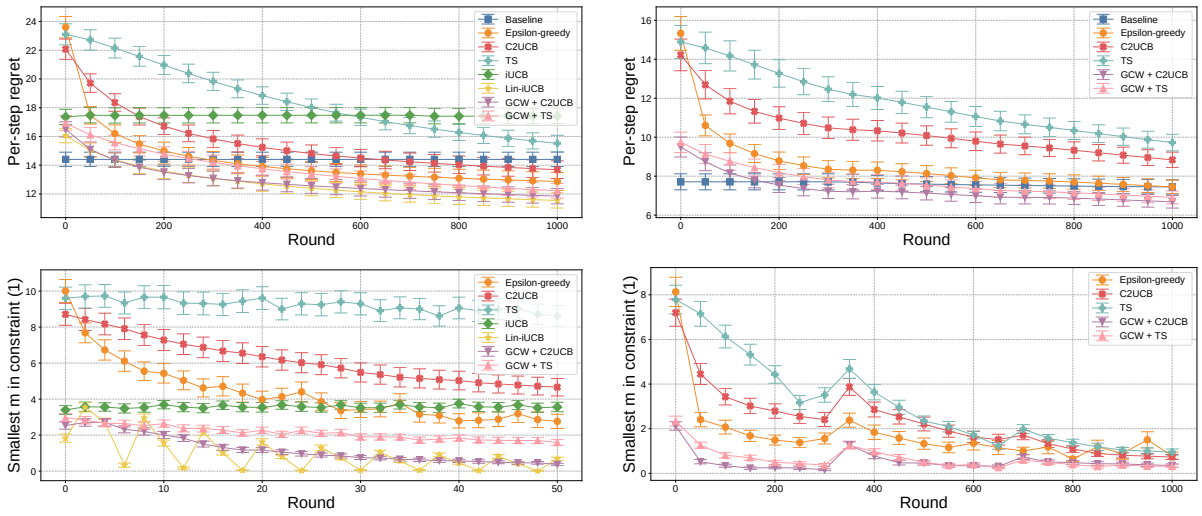


*Figure 1.* Average of the per-step regret (i.e., $R(t)/t$) (*top*) and average of the smallest $m$ satisfying our performance constraint (1) (*bottom*). The error bar represents the standard error. *Left*: the case where $\{x_t(i)\}_{i\in[N]}$, $\mathcal{S}_t$, and $B_t$ are fixed. *Right*: the case where $\{x_t(i)\}_{i\in[N]}$ and $B_t$ change two times.

algorithm, the C$^2$UCB algorithm, and the Thompson sampling in both settings. We additionally compared the iUCB algorithm (Katariya et al., 2019) and its variant (Lin-iUCB algorithm) when they can be applied (i.e., in the first case). The Lin-iUCB used the same reward estimates and confidence widths as those by the proposed algorithm.

Figure 1(*left*) shows the experimental results in the first case. The proposed algorithms outperformed the standard bandit algorithms and the iUCB algorithm. Specifically, the proposed algorithms suffered small regrets in early rounds and outperformed the baseline because of our interleaving technique. The standard bandit algorithms needed large $m$ to satisfy the constraints in early rounds, while these algorithms improved faster than the conservative algorithms by the exploration without the performance constraints. The regret by the Lin-iUCB algorithm was smaller than that by GCW + TS and is compatible with the regret by GCW + C$^2$UCB. However, the Lin-iUCB algorithm was unstable in terms of performance constraints.

Figure 1 (*right*) shows that the proposed algorithms outperformed the existing ones and the baseline in the second case. In addition, GCW + C$^2$UCB outperformed GCW + TS as in the first case. Compared to the results of the first case, all algorithms improved more slowly due to the changing environment. We can see from the numerical results on performance constraints that the quality of recommendations by the bandit algorithms deteriorates temporarily when the environment changes. However, the degradation of the proposed algorithms are smaller than that of other algorithms. This fact suggests that our algorithms are robust to the changes in the given actions.

## 8. Conclusion

We investigated the CCIB problem, which is the CCS problem with stage-wise performance constraints. We proposed the first meta-algorithm for this problem, which preferentially chooses arms with large confidence widths in the baseline arms and those chosen by a given algorithm for the CCS problem. We showed that the proposed algorithm achieves $\widetilde{O}(\min(\sqrt{dkT}, d/\Delta)+dk+\min(k\sqrt{dT/n}, dk/(n\Delta)))$ regret while satisfying the constraints. We also showed a gap-dependent and a gap-independent lower bounds through a reduction to the conservative MAB problem. These lower bounds imply that the proposed algorithm achieves minimax optimal up to logarithmic factors. Our numerical experiments demonstrated that the proposed algorithm improves more efficiently than the iUCB algorithm and outperforms existing algorithms.

There are several open questions about the CCIB problem and its extensions. First, an efficient algorithm for the learner to obtain a bijection that satisfies (3) has not been obtained. Second, it is open whether a similar regret bound can be obtained for the CCIB problem with feasible arms characterized by any matroid. Note that we cannot use the exchange property (3) in this case. Finally, this paper does not consider the problem where the reward of the learner is a non-linear function with respect to the rewards of arms chosen by the learner.

## Acknowledgements

Kei Takemura would like to thank Shinji Ito and Tatsuya Matsuoka for helpful discussions on matroids.

## References


Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 3–11, 1999.

Abeille, M. and Lazaric, A. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the Thirtieth International Conference on International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

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

Garcelon, E., Ghavamzadeh, M., Lazaric, A., and Pirotta, M. Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3962–3969, 2020.

Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *2008 IEEE International Conference on Data Mining*, pp. 263–272, 2008.

Katariya, S., Kveton, B., Wen, Z., and Potluru, V. K. Conservative exploration using interleaving. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 954–963, 2019.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Kazerouni, A., Ghavamzadeh, M., Abbasi-Yadkori, Y., and Van Roy, B. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3910–3919, 2017.

Khezeli, K. and Bitar, E. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10202–10209, 2020.

Korte, B. and Vygen, J. *Combinatorial Optimization: Theory and Algorithms*. Springer, 6th edition edition, 2018.

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

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

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

Moradipari, A., Thrampoulidis, C., and Alizadeh, M. Stage-wise conservative linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11191–11201, 2020.

Qin, L., Chen, S., and Zhu, X. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 461–469, 2014.

Takemura, K. and Ito, S. An arm-wise randomization approach to combinatorial linear semi-bandits. In *2019 IEEE International Conference on Data Mining*, pp. 1318–1323, 2019.

Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9791–9798, 2021.

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.

Wang, Y., Ouyang, H., Wang, C., Chen, J., Asamov, T., and Chang, Y. Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2746–2753, 2017.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. Conservative bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 1254–1262. PMLR, 2016.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pp. 22–32, 2005.

## A. Known Results

Our proofs use the following known results:

**Theorem A.1** (Theorem 13.9 by Korte & Vygen (2018)). *Let $E$ be a finite set and $\mathcal{B} \in 2^E$. $\mathcal{B}$ is the set of bases of some matroid $(E, \mathcal{F})$ if and only if the following holds:*

**(B1)** *$\mathcal{B} \neq \emptyset$;*

**(B2)** *For any $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \setminus B_2$, there exists a $y \in B_2 \setminus B_1$ with $(B_1 \setminus \{x\}) \cup \{y\} \in \mathcal{B}$.*

**Theorem A.2** (Theorem 2 by Abbasi-Yadkori et al. (2011)). *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration, $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $X_t$ is $F_{t-1}$-measurable, $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that $\eta_t$ is $F_t$-measurable. Let $V = \lambda I$ be a positive definite matrix, $V_t = V + \sum_{s\in[t]} X_s X_s^\top$, $Y_t = \sum_{s\in[t]} {\theta^*}^\top X_s + \eta_s$ and $\hat{\theta}_t = V_{t-1}^{-1} Y_t$. Assume for all $t$ that $\eta_t$ is conditionally $R$-sub-Gaussian for some $R > 0$ and $\|\theta^*\|_2 \leq S$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $t \geq 1$,*

$$\|\hat{\theta}_t - \theta^*\|_{V_{t-1}} \leq R\sqrt{2\log\left(\frac{\det(V_{t-1})^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \sqrt{\lambda}S.$$

*Furthermore, if $\|X_t\|_2 \leq L$ for all $t \geq 1$, then with probability at least $1 - \delta$, for all $t \geq 1$,*

$$\|\hat{\theta}_t - \theta^*\|_{V_{t-1}} \leq R\sqrt{d\log\left(\frac{1 + (t-1)L^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S.$$

**Lemma A.3** (Lemma 5 by Takemura et al. (2021)). *Let $\{\{x_t(i)\}_{i\in[k]}\}_{t\in[T]}$ be any sequence such that $x_t(i) \in \mathbb{R}^d$ and $\|x_t(i)\|_2 \leq L$ for all $i \in [k]$ and $t \in [T]$. Let $V_t = \lambda I + \sum_{s\in[t]} \sum_{i\in[k]} x_s(i) x_s(i)^\top$ with $\lambda > 0$. Then, we have*

$$\sum_{t\in[T]} \sum_{i\in[k]} \min\left(\frac{1}{k}, \|x_t(i)\|^2_{V_{t-1}^{-1}}\right) \leq 2d\log(1 + L^2 kT/(d\lambda)).$$

**Lemma A.4** (Lemma 6 by Takemura et al. (2021)). *Let $\{\{x_t(i)\}_{i\in[k]}\}_{t\in[T]}$ be any sequence such that $x_t(i) \in \mathbb{R}^d$ and $\|x_t(i)\|_2 \leq L$ for all $i \in [k]$ and $t \in [T]$. Let $V_t = \lambda I + \sum_{s\in[t]} \sum_{i\in[k]} x_s(i) x_s(i)^\top$ with $\lambda > 0$. Then, we have*

$$\sum_{t\in[T]} \sum_{i\in[k]} \mathbb{1}\left(\|x_t(i)\|_{V_{t-1}^{-1}} > 1/\sqrt{k}\right) \leq 2dk\log(1 + L^2 kT/(d\lambda)).$$

## B. Discussions of Performance Constraints

### B.1. Comparison of Constraint (1) with Constraint (2)

We discuss the fact that the two types of constraints are incomparable. More precisely, we show that each constraint does not include the other constraint.

First, we show that constraint (1) does not include constraint (2).

**Lemma B.1.** *Assume that $m \geq 1$. Then, there exists a set of rewards such that the rewards satisfy the constraint (1) while they do not satisfy the constraint (2) for any $\alpha \in [0, 1)$.*

*Proof.* Suppose that $\mu_t(i) = 0$ for $i \in I_t$. Suppose also that $\mu_t(i) = 1$ for some $i \in B_t$ and $\mu_t(i) = 0$ otherwise. Then, for any bijection $\rho_t : I_t \to B_t$, the constraint (1) is satisfied with $m = 1$. However, since $\sum_{i\in I_t} \mu_t(i) = 0$ and $\sum_{i\in B_t} \mu_t(i) = 1$, the constraint (2) is not satisfied for any $\alpha \in [0, 1)$. □

Next, we show the inverse direction.

**Lemma B.2.** *Assume that $\alpha \in (0, 1]$. Then, there exists a set of rewards such that the rewards satisfy the constraint (2) while they do not satisfy the constraint (1) for any $m < k$.*

*Proof.* We fix $\alpha \in (0, 1]$ arbitrarily. Suppose that $\mu_t(i) = 1 - \alpha$ for $i \in I_t$ and $\mu_t(b) = 1$ for $b \in B_t$. These rewards satisfy the constraint (2) with $\alpha$. Since any reward of arm $b \in B_t$ is strictly larger than any reward of arm $i \in I_t$, we have $\sum_{i\in I_t} \mathbb{1}(\mu_t(i) \geq \mu_t(\rho_t(i))) = 0$ for any bijection $\rho_t : I_t \to B_t$. □

### B.2. Reduction from Constraint (2) to Constraint (1)

We introduce a reduction from constraint (2) to constraint (1). Note that while Assumption 3.10 requires $r_\ell > 0$ due to Assumption 3.9, most of existing studies of conservative bandit problems also require this condition (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020; Khezeli & Bitar, 2020; Moradipari et al., 2020).

**Lemma B.3.** *Under Assumption 3.10, satisfying constraint (1) implies satisfying constraint (2).*

*Proof.* Suppose that the sequence $\{I_t\}_{t\in[T]}$ satisfies the constraints (1). Recall that $\rho_t : I_t \to B_t$ is a bijection. Let $I'_t = \{i \in I_t \mid \mu_t(i) \geq \mu_t(\rho_t(i))\}$ Note that since $\{I_t\}_{t\in[T]}$ satisfies the constraints (1), $|I'_t| \geq k - m$. Rewriting the performance constraints (2), we have

$$\sum_{i\in I'_t}(\mu_t(i) - (1-\alpha)\mu_t(\rho_t(i))) \geq \sum_{i\in I_t\setminus I'_t}((1-\alpha)\mu_t(\rho_t(i)) - \mu_t(i)).$$

Recall that $r_\ell = \min_{t\in[T]}\min_{i\in B_t}\mu_t(i)$. From the definition of $I'_t$ and $r_\ell$, and the fact that $\mu_t(i) \in [0,1]$ for all $i \in [N]$, we have

$$\sum_{i\in I'_t}(\mu_t(i) - (1-\alpha)\mu_t(\rho_t(i))) \geq \alpha\sum_{i\in I'_t}\mu_t(\rho(i)) \geq \alpha(k-m)r_\ell \quad \text{and}$$
$$\sum_{i\in I_t\setminus I'_t}((1-\alpha)\mu_t(\rho_t(i)) - \mu_t(i)) \leq m.$$

Thus, a sufficient condition of the constraints (2) is $(k-m)\alpha r_\ell \geq m$, which is Assumption 3.10. □

## C. Properties of Matroid

### C.1. Strongly Base-Orderable Matroid

First, we introduce the definition of strongly base-orderable matroid and exchangeable set.

**Definition C.1.** A matroid is strongly base-orderable if for any two bases $B_1$ and $B_2$, there is a bijection $f : B_1 \to B_2$ with the property that $(B_1 \setminus X) \cup f(X)$ is a base for any $X \subseteq B_1$.

Then, we show an equivalence between exchangeable set and strongly base-orderable matroid.

**Lemma C.2.** *A set $S$ is exchangeable if and only if the set $S$ is the bases of a strongly base-orderable matroid.*

*Proof.* If the set $S$ is the bases of a strongly base-orderable matroid, we have $S$ is exchangeable by Definition C.1. Suppose that $S$ is exchangeable. We fix $B_1, B_2 \in S$ arbitrarily. Let $\rho : B_1 \to B_2$ be the bijection defined in (3). Since $S$ is exchangeable, we have $(B_1 \setminus \{i\}) \cup \{\rho(i)\} \in S$ for any $i \in B_1 \setminus B_2$. Thus, from Theorem A.1, there exists a matroid such that $S$ is the set of the bases. By Definition 3.6 and Definition C.1, we have the desired result. □

### C.2. Bijection between Two Bases

*Proof of Lemma 5.7.* Let $(E, \mathcal{F})$ be the matroid and $k$ be the rank of the matroid. Let $a^*_1, \ldots, a^*_k$ be a sequence such that $\{a^*_i\}_{i\in[k]} = A^*$ and $r(a^*_i) \leq r(a^*_j)$ for $i \leq j$.

We construct the bijection by induction. We fix $\ell \in [k]$ arbitrarily. Suppose that we have already defined $\rho(a^*_i)$ for all $i < \ell$. Let $A^*_\ell = \{a^*_i\}_{i=\ell,\ldots,k}$ and $A_\ell = A \setminus \{\rho(a^*_i)\}_{i\in[\ell-1]}$. If we have $a^*_\ell \in A_\ell$, we define $\rho(a^*_\ell) = a^*_\ell$. Thus, we consider the other case, i.e., $a^*_\ell \notin A_\ell$. From the augmentation property of matroid, there exists $a \in A_\ell \setminus A^*_{\ell+1}$ such that $A^*_{\ell+1} \cup \{a\} \in \mathcal{F}$. Since $A^* \in \operatorname{argmax}_{I\in\mathcal{A}}\sum_{i\in I} r(i)$, $A^*_\ell$ can be regarded as the output until the $k-\ell+1$-th step of the greedy algorithm. Therefore, we have $r(a^*_\ell) \geq r(a)$ and can define $\rho(a^*_\ell) = a$. □

Note that the construction in our proof of the following lemma is essentially the same to the proof of Lemma 1 of Kveton et al. (2014).

# D. Missing Proofs in Section 5

## D.1. High-Probability Event

For $\delta \in (0,1)$, we define

$$\beta_t(\delta) = \min\left(R\sqrt{2\log\left(N(\pi kt)^2/(3\delta)\right)}, R\sqrt{d\log\left((1+L^2kt/\lambda)/\delta\right)}\right) + M\sqrt{\lambda}$$

for all $t \in [T]$.

*Proof of Lemma 5.3.* We consider two cases: when $\log(N) < d$ and when $\log(N) \geq d$.

First, we consider when $\log(N) < d$. We fix $t \in [T]$ and $i \in [N]$ arbitrarily. From the definition of $\hat{\theta}_t$, we have

$$\begin{aligned}(\hat{\theta}_t - \theta)^\top x_t(i) &= \left(V_{t-1}^{-1}\sum_{s\in[t-1]}\sum_{i\in I_s}(\theta^\top x_s(i) + \eta_s(i))x_s(i) - \theta\right)^\top x_t(i) \\ &= \sum_{s\in[t-1]}\sum_{i\in I_s} x_t(i)^\top V_{t-1}^{-1} x_s(i)\eta_s(i) - \lambda x_t(i)^\top V_{t-1}^{-1}\theta.\end{aligned}$$

For the latter term, we have $\lambda x_t(i)^\top V_{t-1}^{-1}\theta \leq \lambda\|\theta\|_{V_{t-1}^{-1}}\|x_t(i)\|_{V_{t-1}^{-1}} \leq \sqrt{\lambda}M\|x_t(i)\|_{V_{t-1}^{-1}}$. We then bound the former term. Let $\alpha = R\sqrt{2\log(2/\delta')}$ for $\delta' > 0$. Using (a variant of) Azuma's inequality, we obtain

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{s\in[t-1]}\sum_{i\in I_s} x_t(i)^\top V_{t-1}^{-1} x_s(i)\eta_s(i)\right| > \alpha\|x_t(i)\|_{V_{t-1}^{-1}}\right) \\ &\leq 2\exp\left(\frac{\alpha^2\|x_t(i)\|_{V_{t-1}^{-1}}^2}{2R^2\sum_{s\in[t-1]}\sum_{i\in I_s}(x_t(i)^\top V_{t-1}^{-1}x_s(i))^2}\right) \\ &= 2\exp\left(\frac{\alpha^2\|x_t(i)\|_{V_{t-1}^{-1}}^2}{2R^2 x_t(i)^\top V_{t-1}^{-1}(\sum_{s\in[t-1]}\sum_{i\in I_s} x_s(i)x_s(i)^\top)V_{t-1}^{-1}x_t(i)}\right) \\ &\leq 2\exp(-\alpha^2/(2R^2)) \leq \delta'.\end{aligned}$$

Hence, we obtain

$$|(\hat{\theta}_t - \theta)^\top x_t(i)| \leq \left(R\sqrt{2\log(N(\pi kt)^2/(3\delta))} + M\sqrt{\lambda}\right)\|x_t(i)\|_{V_{t-1}^{-1}} \tag{9}$$

with probability at least $1 - 6\delta/(N(\pi kt)^2)$. Taking union bound over the arms in all rounds, we have (9) with probability at least $1-\delta$.

Next, we consider when $\log(N) \geq d$. By the Cauchy-Schwarz inequality, we have

$$|\mu_t(i) - \hat{\theta}_t^\top x_t(i)| \leq \|\theta - \hat{\theta}_t\|_{V_{t-1}}\|x_t(i)\|_{V_{t-1}^{-1}}$$

for all $t \in [T]$ and $i \in [N]$. Then, using Theorem A.2, we have $\|\theta - \hat{\theta}_t\|_{V_{t-1}} \leq R\sqrt{d\log\left(\frac{1+L^2kt/\lambda}{\delta}\right)} + M\sqrt{\lambda}$ for all $t \in [T]$ with probability at least $1-\delta$, which completes the proof. □

## D.2. Analysis for the Recommended Choices

*Proof of Lemma 5.2.* We fix $t \in [T]$ arbitrarily. Let $\tilde{I}_t = I_t \cap \hat{I}_t$ and $\tilde{I}_t^* = \{\rho_t^*(i) \mid i \in \tilde{I}_t\}$.

Let $([N], \mathcal{F}_t)$ be the matroid whose bases are $\mathcal{S}_t$. Let $E_t = \tilde{I}_t \cup \tilde{I}_t^*$ and $k_t = |\tilde{I}_t|$. Then, we can construct a matroid $(E_t, \mathcal{F}_t')$ such that $\mathcal{F}_t' \subseteq \mathcal{F}_t$ and the size of bases of $(E_t, \mathcal{F}_t')$ is $k_t$. Let $\mathcal{S}_t'$ be the bases of $(E_t, \mathcal{F}_t')$. Then, we have

$$\sum_{i\in\tilde{I}_t}(\mu_t(\rho_t^*(i)) - \mu_t(i)) \leq \max_{I\in\mathcal{S}_t'}\sum_{i\in I}\mu_t(i) - \sum_{i\in\tilde{I}_t}\mu_t(i). \tag{10}$$

Recall that $\{r'_t(i)\}_{i\in[N]}$ is the estimation of the rewards by the algorithm $\mathcal{A}$. Since $\hat{I}_t \in \operatorname{argmax}_{I\in\mathcal{S}_t} \sum_{i\in I} r'_t(i)$, we have $\tilde{I}_t \in \operatorname{argmax}_{I\in\mathcal{S}'_t} \sum_{i\in I} r'_t(i)$. Thus, (10) is the regret of the CCS problem in which the feasible set is $\mathcal{S}'_t$. Since the algorithm $\mathcal{A}$ uses only $\{r_t(i)\}_{i\in\tilde{I}_t}$ as feedback in the proposed algorithm, we can bound (10) by the regret upper bound of the algorithm $\mathcal{A}$. □

### D.3. Property of the Bijection Satisfying (3)

**Lemma D.1.** *For any $t \in [T]$ and $I, I' \in \mathcal{S}_t$ such that $I \in \operatorname{argmax}_{J\in\mathcal{S}_t} \sum_{i\in J} r(i)$ for some $r : [N] \to \mathbb{R}$, we have $r(i) \geq r(\rho(i))$ for any $i \in I$, where $\rho : I \to I'$ is the bijection satisfying (3).*

*Proof of Lemma D.1.* We fix $t \in [T]$ and $i \in I_t$ arbitrarily. Let $([N], \mathcal{F})$ be the matroid that corresponds to the constraint $\mathcal{S}_t$. Let $I^{(i)}$ be the set of arms just before the greedy algorithm chooses $i$. By the definition of $\rho$, we have $(I \setminus \{i\}) \cup \{\rho(i)\} \in \mathcal{S}_t$. From the property of matroid, we have $I^{(i)} \cup \{i\} \in \mathcal{F}$ and $I^{(i)} \cup \{\rho(i)\} \in \mathcal{F}$, which finishes the proof. □

### D.4. Properties of Confidence Intervals

To prove Lemma 5.6, we need the following lemma:

**Lemma D.2.** *For any $i \in \hat{I}_t$, if $c_t(i) < \Delta/2$, we have $\mu_t(i) = \mu_t(\hat{\rho}^*_t(i))$.*

*Proof.* Recall that $c_t(i) = \beta_t(\delta)\|x_t(i)\|_{V_{t-1}^{-1}}$. From Lemma D.1, we have $\hat{r}_t(i) \geq \hat{r}_t(\hat{\rho}^*_t(i))$ for all $i \in \hat{I}_t$. Since (6) holds for all $t \in [T]$ and $i \in [N]$, we have

$$\mu_t(\hat{\rho}^*_t(i)) \leq \hat{r}_t(\hat{\rho}^*_t(i)) \leq \hat{r}_t(i) \leq \mu_t(i) + 2c_t(i)$$

for all $i \in \hat{I}_t$. From the assumption of this lemma, we have $\mu_t(\hat{\rho}^*_t(i)) - \mu_t(i) \leq 2c_t(i) < \Delta$. Thus, from the definition of $\Delta$, we obtain $\mu_t(i) = \mu_t(\hat{\rho}^*_t(i))$. □

We are ready to prove Lemma 5.6.

*Proof of Lemma 5.6.* We fix $i \in \bar{I}_t$ arbitrarily. For all $j \in \hat{I}_t$, using Lemma D.2, we have $\mu_t(j) = \mu_t(\hat{\rho}^*_t(j))$. Thus, we have $\mu_t(i) = \mu_t(\bar{\rho}_t(i))$ or $\mu_t(i) - \mu_t(\bar{\rho}_t(i)) \geq \Delta$. If $\mu_t(i) - \mu_t(\bar{\rho}_t(i)) \geq \Delta$, we have

$$\begin{aligned}
\bar{r}_t(\bar{\rho}_t(i)) - \bar{r}_t(i) &= \check{r}_t(\bar{\rho}_t(i)) - \hat{r}_t(i) \\
&\geq (\mu_t(\bar{\rho}_t(i)) - 2c_t(\bar{\rho}_t(i))) - (\mu_t(i) + 2c_t(i)) \\
&\geq \Delta - 2(c_t(i) + c_t(\bar{\rho}_t(i))) \\
&> 0,
\end{aligned}$$

which contradicts the property of $\bar{\rho}_t$ obtained by Lemma D.1. □

### D.5. Number of Rounds in Which Conservative Choices Suffer the Regret

*Proof of Lemma 5.5.* We fix $\ell \in [n]$ arbitrarily. To bound the number of rounds such that $c_t(j^\ell_t) \geq \Delta/4$, we consider a sufficient condition of rounds such that $c_t(j^\ell_t) < \Delta/4$. Since $\beta_T(\delta) \geq \beta_t(\delta)$ for all $t \in [T]$, a sufficient condition of $c_t(j^\ell_t) < \Delta/4$ is that

$$\|x_t(j^\ell_t)\|_{V_{t-1}^{-1}} < \Delta/(4\beta_T(\delta)) = \gamma_\ell/\sqrt{\ell}. \tag{11}$$

Recall that $J^\ell_t$ is defined in line 16 of Algorithm 1. Let $\widetilde{V}_t$ denote $\lambda I + \sum_{s\in[t]} \sum_{j\in J^\ell_s} x_s(j)x_s(j)^\top$. By the definition of $V_t$ and $\widetilde{V}_t$, we have $V_t \succeq \widetilde{V}_t$ for all $t \in [T] \cup \{0\}$. Thus, we obtain

$$\|x_t(i)\|_{V_{t-1}^{-1}} \leq \|x_t(i)\|_{\widetilde{V}_{t-1}^{-1}} \tag{12}$$

for all $t \in [T]$ and $i \in [N]$. Combining (11) and (12), we obtain that a sufficient condition of rounds with $c_t(j_t^\ell) < \Delta/4$ is that $\|x_t(j_t^\ell)\|_{\widetilde{V}_{t-1}^{-1}} < \gamma_\ell/\sqrt{\ell}$. This implies that a necessary condition of rounds with $c_t(j_t^\ell) \geq \Delta/4$ is that $\|x_t(j_t^\ell)\|_{\widetilde{V}_{t-1}^{-1}} \geq \widetilde{\gamma}_\ell$, where $\widetilde{\gamma}_\ell = \min(\gamma_\ell, 1)/\sqrt{\ell}$. Let $T'$ be the number of rounds such that $c_t(j_t^\ell) \geq \Delta/4$. Then, we obtain

$$\widetilde{\gamma}_\ell^2 \ell T' \leq \widetilde{\gamma}_\ell^2 \sum_{t\in[T]} \sum_{j\in J_t^\ell} \mathbb{1}\left(\|x_t(j)\|_{\widetilde{V}_{t-1}^{-1}} \geq \widetilde{\gamma}_\ell\right) \leq \sum_{t\in[T]} \sum_{j\in J_t^\ell} \min\left(\widetilde{\gamma}_\ell^2, \|x_t(j)\|^2_{\widetilde{V}_{t-1}^{-1}}\right) \leq \sum_{t\in[T]} \sum_{j\in J_t^\ell} \min\left(\frac{1}{\ell}, \|x_t(j)\|^2_{\widetilde{V}_{t-1}^{-1}}\right).$$

Applying Lemma A.3 to the last term, we finish the proof. □

### D.6. Regret of an Arm in Conservative Choices

*Proof of Lemma 5.4.* We fix $i \in I_t \setminus \hat{I}_t$ arbitrarily. Recall that $\bar{\rho}_t : \bar{I}_t \to \hat{I}_t$ be the bijection defined in (3). Let $i' = \bar{\rho}_t(i)$. Since (6) holds, we have $\mu_t(i) \geq \hat{r}_t(i) - 2c_t(i)$. Moreover, since $i \in B_t$ and $i' \in \hat{I}_t \setminus B_t$, we have $\hat{r}_t(i) \geq \check{r}_t(i')$ using Lemma D.1. Thus, we obtain

$$\mu_t(i) \geq \hat{r}_t(i) - 2c_t(i) \geq \check{r}_t(i') - 2c_t(i). \tag{13}$$

Recall that $\hat{\rho}_t^* : \hat{I}_t \to I_t^*$ is the bijection defined in (3). From Lemma D.1, we have $\hat{r}_t(i') \geq \hat{r}_t(\hat{\rho}_t^*(i'))$. Thus, we have

$$\mu_t(\hat{\rho}_t^*(i')) \leq \hat{r}_t(\hat{\rho}_t^*(i')) \leq \hat{r}_t(i') = \check{r}_t(i') + 2c_t(i'), \tag{14}$$

where the first inequality is derived from (6).

Recall that $\rho_t^* = \hat{\rho}_t^* \circ \bar{\rho}_t$ for $i \in I_t \setminus \hat{I}_t$. Thus, we have $\hat{\rho}_t^*(i') = \rho_t^*(i)$. Combining (13) and (14), we finish the proof. □

### D.7. Regret Bound for (8)

We can bound the regret as follows:

$$\sum_{t\in[T]} \sum_{j\in J_t^n} R_t(j) = \sum_{t\in[T]} \sum_{j\in \tilde{J}_t^n} R_t(j) + \sum_{t\in[T]} \sum_{j\in \bar{J}_t^n} R_t(j)$$

where

$$\tilde{J}_t^n = \{j \in J_t^n \mid \|x_t(j)\|^2_{\widetilde{V}_{t-1,n}^{-1}} \leq 1/n\} \quad \text{and} \quad \bar{J}_t^n = \{j \in J_t^n \mid \|x_t(j)\|^2_{\widetilde{V}_{t-1,n}^{-1}} > 1/n\}$$

for all $t \in [T]$.

Recall that $\widetilde{V}_{t,n} = \lambda I + \sum_{s\in[t]} \sum_{j\in J_t^n} x_s(j) x_s(j)^\top$. Using Lemma 5.4 and the fact that $V_t \succeq \widetilde{V}_{t,n}$ for all $t \in [T]$, we have

$$\sum_{t\in[T]} \sum_{j\in \tilde{J}_t^n} R_t(j) \leq \sum_{t\in[T]} \sum_{j\in \tilde{J}_t^n} 2\beta_t(\delta) \|x_t(j)\|_{\widetilde{V}_{t-1,n}^{-1}} \quad \text{and}$$

$$\sum_{t\in[T]} \sum_{j\in \tilde{J}_t^n} R_t(j) \leq \sum_{t\in[T]} \sum_{j\in \tilde{J}_t^n} R_t(j)^2/\Delta \leq 2 \sum_{t\in[T]} \sum_{j\in \tilde{J}_t^n} \frac{4\beta_t(\delta)^2}{\Delta} \|x_t(j)\|^2_{\widetilde{V}_{t-1,n}^{-1}}.$$

Thus, we finish the proof by following the same line of our proof of Theorem G.1.

## E. Lower Bounds of Regret

### E.1. Lower Bound when $m = 0$

**Lemma E.1.** *If $m = 0$, for any algorithm, there is an instance of the CCIB problem such that the algorithm does not achieve sub-linear regret while satisfying the performance constraint (1).*

*Proof of Lemma E.1.* We consider two instances of the CCIB problem and show that no algorithm can obtain sub-linear regret while satisfying the constraint for the two instances simultaneously. Let $N = 2$, $d = 2$, $k = 1$ and $m = 0$. Let $x_t(1) = (1, 0)^\top$ and $x_t(2) = (0, 1)^\top$ for all $t \in [T]$, i.e., two-armed bandit problem with the stage-wise performance constraints. Suppose that $r_t(1) = 0.5$ and $B_t = \{1\}$ for all $t \in [T]$. We consider two types of rewards for this setting: One is $r_t(2) = 0$ and the other is $r_t(2) = 1$. In the former case, the baseline is optimal. Thus, to satisfy the constraints, an algorithm must choose the baseline (i.e., the first arm) in all rounds. In the latter case, however, such algorithm does not choose the second arm and then leads to a linear regret. □

### E.2. Lower Bound when $m > 0$

First, we show a gap-independent bound.

**Theorem E.2.** *Consider stochastic conservative MAB problem with $N$ arms and sub-Gaussian noises. Let $R_T^{\mathrm{CMAB}}$ denote the regret of $T$ rounds. Let $i^{\mathrm{base}}$ be the baseline arm and $i_t$ be the arm that are chosen. Suppose that there exist an algorithm and some $\alpha \in \left(0, \frac{1}{4e+2}\right]$ such that the algorithm satisfies $\mathbb{E}\left[\sum_{t\in[T]} \mathbb{1}\left(\bar{r}(i^{\mathrm{base}}) > \bar{r}(i_t)\right)\right] \leq \alpha T$ for any environment, where $\bar{r}(i)$ is the mean reward of arm $i$. Then, there is some environment such that $\mathbb{E}\left[R_T^{\mathrm{CMAB}}\right] > \sqrt{\frac{(N-1)T}{(8e+4)^2\alpha}}$.*

*Proof of Theorem E.2.* Our proof largely follows the proof of Theorem 9 of Wu et al. (2016).

We prove this theorem by contradiction. Thus, we assume that $\mathbb{E}[R_T^{\mathrm{CMAB}}] \leq \sqrt{\frac{(N-1)T}{(8e+4)^2\alpha}}$ for any environment.

We define two types of environments, i.e. the reward distributions of arms and the baseline arm. One is defined as follows:

$$\mu_i = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } i = 1 \\ \mathcal{N}(0, 1) & \text{otherwise} \end{cases},$$

where we abuse $\Delta$ that is defined later. The other type of environments is defined as

$$\mu_i^{(j)} = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } i = 1 \\ \mathcal{N}(2\Delta, 1) & \text{if } i = j \\ \mathcal{N}(0, 1) & \text{otherwise} \end{cases}$$

for $j \in [N] \setminus \{1\}$. Note that these environments have the same sub-optimality gap. Suppose that the baseline arm is the first arm for all environments.

Let $P_\mu(\cdot)$ and $\mathbb{E}_\mu[\cdot]$ denote the probability and the expectation under reward distributions $\mu$, respectively. Let $T_i$ denote the number of times arm $i$ was chosen in the problem. Let $A_i = \{T_i \leq 2\alpha T\}$ and $\Delta = \sqrt{\frac{N-1}{4\alpha T}}$. First, we show $P_\mu(A_i) \geq \frac{1}{2}$ for all $i \in [N] \setminus \{1\}$:

$$P_\mu(A_i) = 1 - P_\mu(T_i > 2\alpha T) \geq 1 - \frac{\mathbb{E}_\mu[T_i]}{2\alpha T} \geq \frac{1}{2},$$

where the first inequality is derived from Markov's inequality, the second inequality follows from the assumption of the algorithm. Next, we show $P_{\mu^{(i)}}(A_i) \leq \frac{1}{4e}$ for all $i \in [N] \setminus \{1\}$:

$$\begin{aligned} P_{\mu^{(i)}}(A_i) &= P_{\mu^{(i)}}(T - T_i \geq T - 2\alpha T) \\ &\leq \frac{\mathbb{E}_{\mu^{(i)}}[T - T_i]}{T - 2\alpha T} \\ &\leq \frac{\mathbb{E}_{\mu^{(i)}}[R_T^{\mathrm{CMAB}}]/\Delta}{T - 2\alpha T} \\ &\leq \frac{1}{(4e+2) - (8e+4)\alpha} \\ &\leq \frac{1}{4e}, \end{aligned}$$

where the first inequality is derived from Markov's inequality and the other inequalities holds due to the definition of the regret and the assumptions. Then, using Lemma 1 of Kaufmann et al. (2016), we have

$$\mathbb{E}_\mu[T_i]\mathrm{KL}(\mu_i, \mu_i^{(i)}) \geq d(P_\mu(A_i), P_{\mu^{(i)}}(A_i)) \geq \frac{1}{2}\log\left(\frac{1}{4(1/(4e))}\right) = \frac{1}{2},$$

where $d(x,y) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$ and $\mathrm{KL}(P,Q)$ is the KL-divergence of $P$ and $Q$. By the definition of $\mu$ and $\mu^{(i)}$, we have $\mathrm{KL}(\mu_i, \mu_i^{(i)}) = 2\Delta^2$. Therefore, we obtain $\mathbb{E}_\mu[T_i] > \frac{1}{4\Delta^2}$ for all $i \in [N] \setminus \{1\}$.

Using this fact, we have $\mathbb{E}_\mu[R_T^{\mathrm{CMAB}}] = \Delta\sum_{i\in[N]\setminus\{1\}}\mathbb{E}_\mu[T_i] > \frac{N-1}{4\Delta} = \Delta\alpha T$, which contradicts the assumption of the algorithm. □

Next, we show a gap-dependent bound.

*Proof of Theorem 6.1.* We prove this theorem by the same line of our proof of Theorem E.2. Thus, we only discuss the difference from that proof.

We assume $\mathbb{E}\left[R_T^{\mathrm{CMAB}}\right] \leq \frac{N-1}{(32e+16)\alpha\Delta}$ for any environment. Let $s$ be the largest number such that $\Delta \leq \sqrt{\frac{N-1}{4\alpha ks}}$. Note that $\mathbb{E}\left[R_{ks}^{\mathrm{CMAB}}\right] \leq \frac{N-1}{(32e+16)\alpha\Delta}$. Let be $A_i = \{ks_i \leq 2\alpha ks\}$, where $ks_i$ is the number of times arm $i$ was chosen until $ks$-th round. We consider the same environments as in Theorem E.2. Then, one can obtain $P_\mu(A_i) \geq 1/2$. Furthermore, since $\Delta \geq \sqrt{\frac{N-1}{4\alpha k(s+1)}}$, we have

$$\begin{aligned}
P_{\mu^{(i)}}(A_i) &= P_{\mu^{(i)}}(ks - ks_i \geq ks - 2\alpha ks) \\
&\leq \frac{\mathbb{E}_{\mu^{(i)}}\left[ks - ks_i\right]}{ks - 2\alpha ks} \\
&\leq \frac{\mathbb{E}_{\mu^{(i)}}\left[R_{ks}^{\mathrm{CMAB}}\right]/\Delta}{ks - 2\alpha ks} \\
&\leq \frac{N-1}{(32e+8)\alpha\Delta^2(ks - 2\alpha ks)} \\
&\leq \frac{s+1}{(8e+4)(s - 2\alpha s)} \\
&\leq \frac{2s}{(8e+4)(s - 2\alpha s)} \\
&= \frac{1}{(4e+2) - (8e+4)\alpha} \\
&\leq \frac{1}{4e},
\end{aligned}$$

where the first inequality is derived from Markov's inequality and the other inequalities holds due to the definition of the regret and the assumptions. Therefore, we obtain $\mathbb{E}_\mu\left[ks_i\right] > 1/(4\Delta^2)$. Finally, we obtain

$$\mathbb{E}_\mu\left[R_{ks}^{\mathrm{CMAB}}\right] > \frac{N-1}{4\Delta} = \Delta\frac{N-1}{4\Delta^2} \geq \Delta\alpha ks,$$

which contradicts the assumption of the algorithm. □

# F. Experiments

## F.1. Environments

In each setting, we set $d = 20$, $T = 1000$, $k = 30$, and $n = 10$. Moreover, any set of $k$ arms can be chosen in each round, i.e., $\mathcal{S}_t = \{I \subseteq [N] \mid |I| = k\}$ for all $t \in [T]$.

*Table 2.* Algorithms in Numerical Experiments

| Algorithm | Parameters |
|---|---|
| $\varepsilon$-greedy | $\varepsilon = 0.05$ and $\lambda = 1$. |
| C$^2$UCB | $\lambda = 1$ and $\delta = 0.05$. |
| Thompson sampling | $\lambda = 1$ and $\delta = 0.05$. |
| iUCB | $\alpha = n/k = 1/3$. |
| Lin-iUCB | $\lambda = 1$, $\delta = 0.05$, and $\alpha = n/k = 1/3$. |
| Proposed | $\lambda = 1$, $\delta = 0.05$ and $n = 10$. |

**Algorithm 2** C$^2$UCB

**Input:** $\lambda > 0$ and $\{\alpha_t\}_{t\in[T]}$ s.t. $\alpha_t > 0$ for all $t \in [T]$.
$V_0 \leftarrow \lambda I$ and $b_0 \leftarrow \mathbf{0}$.
**for** $t = 1, 2, \ldots, T$ **do**
  Observe $\{x_t(i)\}_{i\in[N]}$ and $S_t$.
  $\hat{\theta}_t \leftarrow V_{t-1}^{-1} b_{t-1}$.
  **for** $i \in [N]$ **do**
    $\hat{r}_t(i) \leftarrow \hat{\theta}_t^\top x_t(i) + \alpha_t \sqrt{x_t(i)^\top V_{t-1}^{-1} x_t(i)}$.
  **end for**
  Play a set of arms $I_t \in \operatorname{argmax}_{I\in\mathcal{S}_t} \sum_{i\in I} \hat{r}_t(i)$ and observe rewards $\{r_t(i)\}_{i\in I_t}$.
  $V_t \leftarrow V_{t-1} + \sum_{i\in I_t} x_t(i) x_t(i)^\top$ and $b_t \leftarrow b_{t-1} + \sum_{i\in I_t} r_t(i) x_t(i)$.
**end for**

We use the Book-Crossing dataset (Ziegler et al., 2005), which consists of ratings expressed on a scale from 0 to 10 for books. We extracted users that rated at least 100 items and items that are rated by at least 3 users in the extracted users, which obtains $37,157$ ratings about $7,068$ items by 490 users.

We consider the cold start problem as in Garcelon et al. (2020). An algorithm aims to learn the preference of a new user. We obtained the item vectors from the ratings by the weighted regularized matrix factorization (Hu et al., 2008) and used the item vectors as the feature vectors. The baseline is the recommendation by the matrix factorization. A user is randomly selected at the beginning of a simulation. The mean rewards are the selected user's ratings that are normalized in $[0, 1]$. The noises of the rewards follow $\mathcal{N}(0, 0.1^2)$. We ran 100 simulations.

### F.2. Algorithms

We compare the proposed algorithm with the algorithms whose parameters are described in Table 2 and the baseline. The $\varepsilon$-greedy algorithm has two ways to estimate the rewards of given arms: One is to use random values, and the other is to use the ridge regression where $\lambda$ is the parameter of the regularization. The algorithm chooses the former way with probability $\varepsilon$ and the latter way otherwise. Then, it plays a feasible set of arms that maximizes the sum of the estimated rewards. We set $\alpha$ of the iUCB algorithms[5] such that the iUCB and the proposed algorithms have the same condition for exploration, i.e., $\alpha = n/k$. Note that $\alpha$ for the iUCB algorithm and $\alpha$ in (2) are different parameters.

## G. C$^2$UCB Algorithm for Matroid Constraint

We consider the CCS problem defined in Takemura et al. (2021). Note that this problem assumes that $|\mu_t(i)| \leq B$ for some parameter $B$ but does not assume that $\mu_t(i) \in [0, 1]$. The CCS problem coincides with the CCIB problem without the performance constraints except the above assumption.

We show that the C$^2$UCB algorithm (Algorithm 2) is minimax optimal up to logarithmic factors for the CCS problem with *any* matroid constraint. We emphasize that our theorem is a generalization of Theorem 3 by Takemura et al. (2021) and that our proof is much simpler than that of the theorem by Takemura et al. (2021). While the CCS problem assumes that the

[5] Strictly speaking, we implemented the iUCB2 algorithm, which does not need to know the rewards of the baseline arms in advance.

number $k$ of arms chosen in a round is fixed over the rounds, the C$^2$UCB algorithm works and achieves the same regret bound if the number of arms chosen in a round is bounded by $k$ from above.

Let $\rho_t^* : I_t \to I_t^*$ be the bijection defined by Lemma 5.7 with $\{\hat{r}_t(i)\}_{i\in[N]}$. Then, we define our sub-optimality gap as follows:

$$\Delta_t(i,j) = \mu_t(j) - \mu_t(i) \quad \text{and} \quad \Delta = \min_{t\in[T]} \min_{i\in I_t : \Delta_t(i,\rho_t^*(i))>0} \Delta_t(i,\rho_t^*(i)).$$

Using this sub-optimality gap, we have the following regret bound:

**Theorem G.1** (A generalization of Theorem 3 by Takemura et al. (2021))**.** *Let $\kappa = \min(\log(N), d)$, $\nu = \log(kT)$ and $\iota = \min(\log(NkT/\delta), d\log(kT/\delta))$. Assume that $\mathcal{S}_t$ is a set of bases of a matroid for all $t \in [T]$. Then, if $\alpha_t = \beta_t(\delta)$ and $\lambda = (R/M)^2\kappa$, the C$^2$UCB algorithm has the following regret bound with probability $1-\delta$*

$$R(T) = O\left(\min\left(R\sqrt{dkT\nu\iota}, \frac{R^2 d\nu\iota}{\Delta}\right) + Bdk\nu\right).$$

*Proof.* Let $J_t = \{i \in I_t \mid \|x_t(i)\|_{V_{t-1}^{-1}} > 1/\sqrt{k}\}$ and $J_t^* = \{\rho_t^*(i) \mid i \in J_t\}$. Using these notations, the regret is rewritten as

$$R(T) = \sum_{t\in[T]} \left(\sum_{i\in I_t^*\setminus J_t^*} \mu_t(i) - \sum_{i\in I_t\setminus J_t} \mu_t(i)\right) + \sum_{t\in[T]} \left(\sum_{i\in J_t^*} \mu_t(i) - \sum_{i\in J_t} \mu_t(i)\right).$$

We bound these terms separately.

First, we consider the former term. By the definition of the bijection $\rho_t^*$, we have $\hat{r}_t(i) \geq \hat{r}_t(\rho_t^*(i))$ for all $i \in I_t$. Using Lemma 5.3, we have $\hat{r}_t(i) \geq \mu_t(\rho_t^*(i))$ and $\hat{r}_t(i) \leq \mu_t(i) + 2c_t(i)$ for all $i \in I_t$, where $c_t(i) = \beta_t(\delta)\|x_t(i)\|_{V_{t-1}^{-1}}$. Therefore, we obtain

$$\begin{aligned}
\sum_{t\in[T]} \left(\sum_{i\in I_t^*\setminus J_t^*} \mu_t(i) - \sum_{i\in I_t\setminus J_t} \mu_t(i)\right) &\leq \sum_{t\in[T]} \sum_{i\in I_t\setminus J_t} 2c_t(i) \quad \text{and} \\
\sum_{t\in[T]} \left(\sum_{i\in I_t^*\setminus J_t^*} \mu_t(i) - \sum_{i\in I_t\setminus J_t} \mu_t(i)\right) &\leq \sum_{t\in[T]} \left(\sum_{i\in I_t\setminus J_t} (\mu_t(\rho_t^*(i)) - \mu_t(i))^2\right)/\Delta \\
&\leq \sum_{t\in[T]} \sum_{i\in I_t\setminus J_t} 4c_t(i)^2/\Delta
\end{aligned}$$

Since $\beta_T(\delta) \geq \beta_t(\delta)$ for all $t \in [T]$, using Lemma A.3, we have

$$\begin{aligned}
\sum_{t\in[T]} \sum_{i\in I_t\setminus J_t} c_t(i) \leq \beta_T(\delta) \sum_{t\in[T]} \sum_{i\in I_t\setminus J_t} \|x_t(i)\|_{V_{t-1}^{-1}} &\leq \beta_T(\delta)\sqrt{kT \sum_{t\in[T]} \sum_{i\in I_t\setminus J_t} \|x_t(i)\|^2_{V_{t-1}^{-1}}} \\
&= O(R\sqrt{dkT\iota}) \quad \text{and} \\
\sum_{t\in[T]} \sum_{i\in I_t\setminus J_t} c_t(i)^2/\Delta &= O(R^2 d\iota/\Delta).
\end{aligned}$$

Next, we bound the latter term. From Lemma A.4, we have

$$\sum_{t\in[T]} |J_t| = \widetilde{O}(dk).$$

Since $|\mu_t(i)| \leq B$, we have

$$\sum_{t\in[T]} \left(\sum_{i\in J_t^*} \mu_t(i) - \sum_{i\in J_t} \mu_t(i)\right) \leq 2B \sum_{t\in[T]} |J_t| = O(Bdk\nu).$$

□

**Algorithm 3** Thompson sampling

**Input:** $\lambda > 0$ and $\{v_t\}_{t\in[T]}$ for all $t \in [T]$.
$V_0 \leftarrow \lambda I$ and $b_0 \leftarrow \mathbf{0}$.
**for** $t = 1, 2, \ldots, T$ **do**
  Observe $\{x_t(i)\}_{i\in[N]}$ and $S_t$.
  $\hat{\theta}_t \leftarrow V_{t-1}^{-1} b_{t-1}$.
  Sample $\tilde{\theta}_t$ from $\mathcal{N}\left(\hat{\theta}_t, v_t^2 V_{t-1}^{-1}\right)$
  **for** $i \in [N]$ **do**
    $\tilde{r}_t(i) \leftarrow \tilde{\theta}_t^\top x_t(i)$.
  **end for**
  Play a set of arms $I_t \in \operatorname{argmax}_{I\in\mathcal{S}_t} \sum_{i\in I} \tilde{r}_t(i)$ and observe rewards $\{r_t(i)\}_{i\in I_t}$.
  $V_t \leftarrow V_{t-1} + \sum_{i\in I_t} x_t(i) x_t(i)^\top$ and $b_t \leftarrow b_{t-1} + \sum_{i\in I_t} r_t(i) x_t(i)$.
**end for**

## H. Thompson Sampling for Matroid Constraint

We show that the Thompson sampling Algorithm 3 for the CCS problem achieves a near-optimal regret bound. Note that, similar to the C$^2$UCB algorithm, in the CCS problem where the number of arms chosen in $t$-th round depends on $t$, the Thompson sampling achieves the same regret bound if the number of arms chosen in a round is bounded by $k$ from above.

**Theorem H.1.** *Let $\kappa = \min(\log(N), d)$, $\nu = \log(kT)$ and $\iota' = \min(\log(NkT^2/\delta), d\log(kT^2/\delta))\log(dT/\delta)$. Assume that $\mathcal{S}_t$ is a set of bases of a matroid for all $t \in [T]$. Then, if $v_t = \beta_t(\delta/(4T))$ and $\lambda = L/(dM)$, the Thompson sampling has the following regret bound with probability $1-\delta$*

$$R(T) = O\left((R\sqrt{d\kappa} + \sqrt{ML})\sqrt{dkT\nu\iota'} + Bdk\nu\right).$$

*Proof.* Let $\rho_t^* : I_t \to I_t^*$ be the bijection defined by Lemma 5.7 with $\{\tilde{r}_t(i)\}_{i\in[N]}$. Let $J_t = \{i \in I_t \mid \|x_t(i)\|_{V_{t-1}^{-1}} > 1/\sqrt{k}\}$ and $J_t^* = \{\rho_t^*(i) \mid i \in J_t\}$. Using these notations, the regret is rewritten as

$$\begin{aligned} R(T) = &\sum_{t\in[T]} \left( \sum_{i\in I_t^* \setminus J_t^*} \mu_t(i) - \sum_{i\in I_t \setminus J_t} \mu_t(i) \right) \\ &+ \sum_{t\in[T]} \left( \sum_{i\in J_t^*} \mu_t(i) - \sum_{i\in J_t} \mu_t(i) \right). \end{aligned} \tag{15}$$

We bound these terms separately.

First, we bound the latter term. By the same discussion in the proof of Theorem G.1, we have

$$\sum_{t\in[T]} \left( \sum_{i\in J_t^*} \mu_t(i) - \sum_{i\in J_t} \mu_t(i) \right) \le 2B \sum_{t\in[T]} |J_t| = O(Bdk\nu).$$

Next, we consider the former term. If $\mu_t(i) \ge \mu_t(\rho_t^*(i))$, the term $\mu_t(\rho_t^*(i)) - \mu_t(i)$ does not contribute to the regret. Thus, we can assume

$$\forall i \in I_t \setminus J_t,\ \mu_t(i) \le \mu_t(\rho_t^*(i)) \tag{16}$$

without loss of generality.

We follow a similar line of proof for the Thompson sampling for the contextual linear bandit problem by Abeille & Lazaric (2017). As in Abeille & Lazaric (2017), we define the following events for all $t \in [T]$:

$$\begin{aligned} \hat{E}_t &= \{\forall s \le t, \|\hat{\theta}_s - \theta\|_{V_{s-1}} \le \beta_s(\delta')\} \quad \text{and} \\ \tilde{E}_t &= \{\forall s \le t, \|\tilde{\theta}_s - \hat{\theta}_s\|_{V_{s-1}} \le \gamma_s(\delta')\}, \end{aligned}$$

where $\delta' = \delta/(4T)$ and $\gamma_t(\delta) = \sqrt{2d\log(2d/\delta)}\beta_t(\delta)$. From a similar argument of Lemma 1 by Abeille & Lazaric (2017), we have $P(\hat{E}_T \cap \tilde{E}_T) \geq 1 - \delta/2$. In the remaining of this proof, we consider the case under the event $\hat{E}_T \cap \tilde{E}_T$.

Then, we can decompose the right-hand side of (15) as follows:

$$\sum_{t\in[T]} \left( \sum_{i\in I_t^* \setminus J_t^*} \mu_t(i) - \sum_{i\in I_t \setminus J_t} \mu_t(i) \right) = \sum_{t\in[T]} \left( \sum_{i\in I_t \setminus J_t} \theta^\top x_t(\rho_t^*(i)) - \tilde{\theta}_t^\top x_t(i) \right) \tag{17}$$

$$+ \sum_{t\in[T]} \left( \sum_{i\in I_t \setminus J_t} \tilde{\theta}_t^\top x_t(i) - \theta^\top x_t(i) \right). \tag{18}$$

We can bound (18) as

$$\sum_{t\in[T]} \left( \sum_{i\in I_t \setminus J_t} \tilde{\theta}_t^\top x_t(i) - \theta^\top x_t(i) \right) \leq \sum_{t\in[T]} \sum_{i\in I_t \setminus J_t} (\gamma_t(\delta') + \beta_t(\delta')) \|x_t(i)\|_{V_{t-1}^{-1}} = O\left( (R\sqrt{\kappa} + M\sqrt{\lambda}) d\sqrt{kT\nu\iota'} \right).$$

To bound (17), we introduce convex functions $J_{t,i}(u) = \max_{j\in C_{t,i}} u^\top x_t(j)$ for some $C_{t,i} \subseteq [N]$ such that $J_{t,i}(\tilde{\theta}_t) = \tilde{\theta}_t^\top x_t(i)$ and $J_{t,i}(\theta) = \theta^\top x_t(\rho_t^*(i))$. If we construct such convex functions, by the same line of the proof by Abeille & Lazaric (2017), we can complete the proof. In fact, we have

$$\sum_{t\in[T]} \left( \sum_{i\in I_t \setminus J_t} \theta^\top x_t(\rho_t^*(i)) - \tilde{\theta}_t^\top x_t(i) \right) = O\left( (R\sqrt{\kappa} + M\sqrt{\lambda}) d\sqrt{kT\nu\iota'} + \frac{L}{\sqrt{\lambda}} \sqrt{kT\log(1/\delta)} \right)$$

with probability at least $1 - \delta/2$.

We now construct $J_{t,i}(u)$. By the construction of $\rho_t^*$, we have $\tilde{\theta}_t^\top x_t(i) \geq \tilde{\theta}_t^\top x_t(\rho_t^*(i))$. On the other hand, we have $\theta^\top x_t(i) \leq \theta^\top x_t(\rho_t^*(i))$ by (16). Therefore, if we define $C_{t,i} = \{i, \rho_t^*(i)\}$, we have $J_{t,i}(\tilde{\theta}_t) = \tilde{\theta}_t^\top x_t(i)$ and $J_{t,i}(\theta) = \theta^\top x_t(\rho_t^*(i))$. □