The Hardness Analysis of Thompson Sampling for Combinatorial Semi-bandits with Greedy Oracle

Fang Kong\textsuperscript{1}  \quad Yueran Yang\textsuperscript{1}  \quad Wei Chen\textsuperscript{2}  \quad Shuai Li\textsuperscript{1}*  
\textsuperscript{1}Shanghai Jiao Tong University  \quad \textsuperscript{2}Microsoft Research  
{fangkong, yangyr99, shuaili8}@sjtu.edu.cn  \quad weic@microsoft.com

Abstract

Thompson sampling (TS) has attracted a lot of interest in the bandit area. It was introduced in the 1930s but has not been theoretically proven until recent years. All of its analysis in the combinatorial multi-armed bandit (CMAB) setting requires an exact oracle to provide optimal solutions with any input. However, such an oracle is usually not feasible since many combinatorial optimization problems are NP-hard and only approximation oracles are available. An example \cite{30} has shown the failure of TS to learn with an approximation oracle. However, this oracle is uncommon and is designed only for a specific problem instance. It is still an open question whether the convergence analysis of TS can be extended beyond the exact oracle in CMAB. In this paper, we study this question under the greedy oracle, which is a common (approximation) oracle with theoretical guarantees to solve many (offline) combinatorial optimization problems. We provide a problem-dependent regret lower bound of order $\Omega(\log T/\Delta^2)$ to quantify the hardness of TS to solve CMAB problems with greedy oracle, where $T$ is the time horizon and $\Delta$ is some reward gap. We also provide an almost matching regret upper bound. These are the first theoretical results for TS to solve CMAB with a common approximation oracle and break the misconception that TS cannot work with approximation oracles.

1 Introduction

Stochastic multi-armed bandit (MAB) problem \cite{20, 5, 3} is a classical online learning framework. It has been extensively studied in the literature and has a wide range of applications \cite{13, 12, 20}. The problem is modeled by a $T$-round game between the learning agent and the environment. The environment contains an arm set and each arm is associated with a reward distribution. At each round $t$, the agent first selects an arm, while the environment generates a random reward for each arm from its reward distribution. The agent then obtains the reward of the selected arm. The objective of the agent is to accumulate as many expected rewards over $T$ rounds, or equivalent to minimizing the cumulative expected regret, which is defined as the cumulative difference between the expected reward of the optimal arm and the selected arms over $T$ rounds. To achieve this long-horizon goal, the learning agent has to face the dilemma of exploration and exploitation in each round. The former aims to try arms that have not been observed enough times to get a potentially higher reward, the latter focuses on the arm with the best observed performance so far to maintain a high profit. How to balance the tradeoff between exploration and exploitation is the main focus of the MAB algorithms.

One of the most popular bandit algorithms is the upper confidence bound (UCB) algorithm \cite{5}. The algorithm aims to construct confidence sets for unknown expected rewards and selects arms according to their highest upper confidence bounds. The UCB-type algorithms have been widely studied and provided theoretical guarantees with regret upper bound of order $O(\log T/\Delta)$, where $\Delta$ is the
minimum gap between the expected reward of the optimal arm and any suboptimal arms. Thompson sampling (TS) [3, 26] is another popular method to solve MAB problems. It is a randomized algorithm based on the Bayesian idea, which maintains an iteratively updated posterior distribution for each arm and chooses arms according to their probabilities of being the best one. The TS algorithm was introduced in the 1930s [28], but its theoretical analysis is open until recent years [15, 2]. It was proven that the regret upper bound of the TS algorithm is of the same order of $O(\log T/\Delta)$ in MAB problems [3]. Benefited from the advantages of easier implementation and better empirical performance compared to UCB, the TS-type algorithms attract more attentions in recent years.

Despite its importance, the MAB framework may fail to model many real applications since the agent’s action is usually not a single arm but a combination of several arms. This motivates the study on the combinatorial MAB (CMAB) problem [11, 9, 17, 31, 7, 29, 30, 24]. In the CMAB framework, the agent selects a combination of base arms as an action to play in each round. All outcomes of these selected arms are then revealed to the agent, which is called semi-bandit feedback and is widely studied in the literature [7, 29, 32, 30, 14, 24]. Such CMAB framework can cover many real application scenarios including probabilistic maximum coverage (PMC) [6], online influence maximization (OIM) [29], multiple-play MAB (MP-MAB) [17] and minimum spanning tree (MST).

TS-type algorithms have recently attracted a lot of interest in CMAB problems [17, 30, 14, 24]. All of these works need an exact oracle to provide the optimal solution with sampled parameters as input in each round. However, such oracles are usually not feasible since many offline combinatorial optimization problems, such as the offline problem of PMC and OIM [16], are NP-hard and only approximation oracles are available. With an example [30] illustrating the non-convergent regret of TS with an artificial approximation oracle designed for a specific problem instance, whether TS can work well in CMAB problems with common approximation oracles is still an open problem.

One of the most common oracles for offline combinatorial optimization problems with a theoretical guarantee is the greedy algorithm. It sequentially finds the current optimal arm, or the optimal collection of multiple arms according to the structural correlations, to maximize the current total expected reward. When the termination condition is met, it will return the set of all arms found in previous steps as the solution. The termination condition is usually formulated by a constraint on the number of steps. For example, in the PMC, OIM, and MP-MAB problems, such a process is limited to continue $K$ steps. The greedy algorithm can provide approximate solutions for offline problems of PMC [8] and OIM [16], and exact optimal solutions for offline problems of MP-MAB [17] and MST [18]. In general, as long as the expected reward in a problem satisfies the monotonicity and submodularity on the action set, the greedy algorithm serves as an offline oracle to provide an approximate solution [23].

In this paper, we first formulate the CMAB problems with greedy oracle, which is general enough and covers PMC, MP-MAB, and many other problems. In this framework, the objective of the learning agent is to minimize the cumulative greedy regret, defined as the cumulative difference between the expected reward of the selected action and that of the greedy’s solution in the real environment. Focusing on a specific PMC problem instance, we derive the hardness analysis of the TS algorithm with the greedy oracle to solve CMAB problems. Due to the mismatch between the estimation gaps that need to be eliminated by exploration and the actual regret the algorithm needs to pay for each such exploration, the TS algorithm with greedy oracle in this CMAB problem cannot achieve as good theoretical performance as previous MAB algorithms. A problem-dependent regret lower bound of order $\Omega(\log T/\Delta^2)$ is provided to illustrate such hardness, where $T$ is the time horizon and $\Delta$ is some reward gap. By carefully exploiting the property of the greedy oracle, we also provide an almost matching problem-dependent regret upper bound, which is tight on that PMC problem instance and also recovers the main order of TS when solving MAB problems [3]. These results are the first theoretical results of TS with approximation oracle to solve CMAB problems, which show that the linear regret example in [30] does not hold for every approximation oracle.

## 2 Related Work

CMAB problems have been widely studied in the literature [11, 6, 7, 29, 30, 24, 32, 21, 14]. Here we mainly focus on the most relevant works. Gai et al. [11] first study the CMAB problem with linear reward, where the reward of an action is linear in the reward of base arms included in it. They introduce a UCB-type algorithm to solve such problems and allow approximation algorithms to
serve as offline oracles. Later, Chen et al. [6, 7, 29] generalize this framework by considering a larger class of rewards and the case with probabilistically triggered arms (CMAB-T). This framework only assumes the expected reward satisfies the monotonicity and Lipschitz condition on the mean vector of base arms. The combinatorial UCB (CUCB) algorithm is proposed to solve such general CMAB problems, which works with any offline oracle with approximation guarantees. When only approximation oracles are available, the goal of the algorithm is relaxed to minimize the cumulative approximation regret, which is defined as the difference between the expected reward of the selected action and that of the scaled optimal solution. The CUCB algorithm achieves the regret upper bound of order $O(\log T/\Delta_{\min})$ [29], where $\Delta_{\min}$ is the minimum reward gap from the scaled optimal solution over all suboptimal actions.

Compared with UCB-type algorithms which need to compute upper confidence bounds for unknown means of base arms [6, 7, 29], TS-type algorithms do not require the reward function to satisfy the monotonicity on the mean vector of base arms. Benefited from this and other advantages of easier implementation and better practical performances, TS-type algorithms have recently attracted a lot of interest in CMAB problems. Komiyama et al. [17] consider using the TS algorithm to solve the MP-MAB problem, where the agent needs to select $K$ from $m$ arms to maximize the sum of rewards over these selected $K$ arms. They provide an optimal regret upper bound of order $O(\log T/\Delta_{K,K+1})$ for the TS algorithm to solve this problem, where $\Delta_{K,K+1}$ is the reward gap between the $K$-th and $(K+1)$-th optimal arm. Later, Wang and Chen [30] consider using TS to solve more general CMAB problems where only the Lipschitz condition is assumed to be satisfied. Their regret upper bound of order $O(\log T/\Delta_{\min})$ matches the main order of the CUCB [29] in the same setting. The coefficient of this upper bound was recently improved by Perrault et al. [24], who study the same CMAB setting. Huyuk et al. [14] extend the analysis of [30] and consider using the TS algorithm to solve the problem of CMAB-T. However, the current regret upper bound is $O(1/p^*)$ worse than CUCB [29], where $p^*$ is the minimum triggering probability.

All of the above TS-based works need an exact oracle to provide the optimal solution with sampled parameters in each round. However, the exact oracles are usually not feasible as many combinatorial optimization problems, such as the offline problem of OIM and PMC, are NP-hard [16]. Wang and Chen [30] have constructed a problem instance and designed an approximation offline oracle for this problem instance. The analysis has shown that the TS algorithm suffers the linear regret of order $O(T)$ when working with such an approximation oracle. However, this oracle is uncommon and artificial, thus cannot represent the performance of TS when working with common approximation oracles. It is still a significant open problem that whether TS can perform well with approximation oracles.

The greedy algorithm is one of the most important methods with approximation guarantees to solve combinatorial optimization problems. When the mean vector is known beforehand, we call the problem of finding the action with the best expected reward as offline problem. Using the greedy algorithm to solve offline combinatorial problems has been studied for decades, including the problem of shortest spanning subtree [18], shortest connection networks [25], set coverage [8], influence maximization [16], and general submodular optimization [23]. There is also a line of studies considering using greedy to solve online problems [4, 19, 10, 27, 22], some of them require the exact reward function forms as prior knowledge. Among these works the most related to ours is [22], both aiming to solve a general class of online problems. Lin et al. [22] consider using the online greedy strategy to make decisions based on UCB-type estimators. The algorithm sequentially selects a unit to maximize the current expected reward until no feasible unit can be selected in each round. In their setting, a unit conditioned on a set of previously selected units is regarded as an arm and the marginal reward of selecting this unit is the expected reward of this arm. Since the number of combinations of units is usually exponentially large, there is an exponential number of arms to explore, making the algorithm pay exponential memory cost in practical applications. Based on this framework, the algorithm needs to observe the marginal reward after the decision of each step to update the estimate on the arm. However, such observation may be not available as many combinatorial problems treat the action composed of several units as a whole and select them together. Compared to this work, our framework only needs polynomial memory cost and does not require the observation of the marginal reward.

In this paper, we study the problem of CMAB with the common (approximation) greedy oracle and hope to answer the question of whether the TS algorithm can work well in this setting.
3 Setting

The combinatorial multi-armed bandit (CMAB) problem is formulated by a $T$-round learning game between the learning agent and the environment. The environment contains $m$ base arms and the arm set is denoted by $[m] = \{1, 2, \ldots, m\}$. Each arm $i \in [m]$ is associated with a distribution $D_i$ on $[0, 1]$. We consider a combinatorial setting where the agent can select several base arms at a time. In many applications, different arms usually have structural correlations in the selection decision of the agent. For example, in the PMC problem, base arms (edges) starting from the same node must be selected together. Thus the base arm set $[m]$ can be further divided into $n$ units, with each unit containing several base arms and a unit of arms will be selected together. Let $\mathcal{U}$ be the collection of all units and $|s|$ be the number of base arms contained in unit $s$ for any $s \in \mathcal{U}$.

In each round $t = 1, 2, \ldots$, the learning agent selects an action $S_t \subseteq \mathcal{S} = \{S \subseteq \mathcal{U} : |S| = K\}$ to play. Here $\mathcal{S}$ is the set of all candidate actions containing $K$ units. For any action $S$, denote $\cup S = \{i \in s \text{ for some } s \in S\}$ as the set of base arms that belong to units contained in $S$. The environment then draws a random output of all base arms $X_t = (X_{t, 1}, X_{t, 2}, \ldots, X_{t, m})$ from the distribution $D = D_1 \times D_2 \times \ldots \times D_m$. For any $t$, $X_{t,i}$ is independent and identically distributed on $D_i$ with expectation $\mu_i$, for any base arm $i$. Let $\mu = (\mu_i)_{i \in [m]}$ be the mean vector. We study the semi-bandit feedback [30, 7, 29] where the agent can observe feedback $Q_t = \{(i, X_{t,i}) \mid i \in \cup S_t\}$, namely the output of all base arms in units contained in $S_t$. Denote $\mathcal{H}_t = \{(S_t, Q_t) : 1 \leq \tau < t\}$ as the history of observations at time $t$. The agent finally obtains a corresponding reward $R_t = R(S_t, X_t)$ in this round, which is a function of action $S_t$ and output $X_t$. We assume the expected reward satisfies the following two assumptions, which are standard in CMAB works [7, 30, 14, 24, 29].

**Assumption 1.** The expected reward of an action $S$ only depends on $S$ and the mean vector $\mu$. That is to say, there exists a function $r$ such that $E[R_t] = E_{X_t \sim D}(R(S_t, X_t)) = r(S_t, \mu)$.

**Assumption 2.** (Lipschitz continuity) There exists a constant $B$ such that for any action $S$ and mean vectors $\mu, \mu'$, the reward of $S$ under $\mu$ and $\mu'$ satisfies

$$|r(S, \mu) - r(S, \mu')| \leq B \sum_{i \in \cup S} |\mu_i - \mu'_i|.$$  (1)

When the mean vector $\mu$ is known beforehand, finding the optimal action containing $K$ units is called the offline problem. However, the offline problems are usually NP-hard and enumerating all actions to find the best one is not feasible as the number of actions is exponentially large. The Greedy algorithm (presented in Algorithm 1) is a common method to solve such offline problems, which is simple to implement and can provide approximate solutions for OIM [16] and PMC [8], and exact solutions for MP-MAB[17]. More specifically, as long as the reward function satisfies monotonicity and submodularity on the action set, the Greedy algorithm can provide solutions with approximate guarantees [23]. Moreover, the Greedy algorithm is also popular to serve as a heuristic method in real applications and has good practical performance even without a theoretical guarantee.

**Algorithm 1 Greedy algorithm**

1: Input: base arm set $[m]$ and mean vector $\mu = (\mu_i)_{i \in [m]}$, unit set $\mathcal{U}$, action size $K$
2: Initialize: $S_g = \emptyset$
3: for $k = 1, 2, \ldots, K$ do
4:     $s_k = \text{argmax}_{s \in \mathcal{U} \setminus S_g} r(S_g \cup \{s\}, \mu)$
5:     $S_g = S_g \cup \{s_k\}$
6: end for
7: Output: $S_g$

We mainly study the CMAB problem with the Greedy oracle. With input $\mu = (\mu_i)_{i \in [m]}$, it sequentially selects $K$ units to maximize the current expected reward. To simplify, we assume the Greedy’s solution $S_g(\mu)$, abbreviated as $S_g$, is unique, or equivalently the optimal unit in each step $t$ (Line 4 in Algorithm 1) is unique. The general case with multiple solutions can also be solved and would be discussed later. The objective of the learning agent is to maximize the cumulative expected reward over $T$ rounds, or equivalently to minimizing the cumulative expected regret with respect to the
Greedy’s solution $S_g$, which we call cumulative greedy regret [22] defined by

$$R_g(T) = \mathbb{E} \left[ \sum_{t=1}^{T} \max \{ r(S_g, \mu) - r(S_t, \mu), 0 \} \right], \quad (2)$$

where the expectation is taken from the randomness in observations and the online algorithm.

We call Greedy an $\alpha$-approximate oracle if $r(S_g(\mu'), \mu') \geq \alpha \cdot r(S^*(\mu'), \mu')$ for any input $\mu'$, where $S^*(\mu')$ is the optimal action under $\mu'$. Note when Greedy is $\alpha$-approximate, the upper bound for greedy regret also implies the upper bound for the $\alpha$-approximate regret defined by the cumulative distance between scaled optimal reward $\alpha \cdot r(S^*(\mu), \mu)$ and $r(S_t, \mu)$ over $T$ rounds. The approximate regret is adopted in previous CMAB works based on UCB-type algorithms [7, 29, 32, 21]. It is much weaker than greedy regret as it relaxes the requirements for online algorithms and only needs them to return solutions satisfying the relaxed approximation ratio. We discuss more on challenges in analyzing the $\alpha$-approximate regret with TS-type algorithms in Section 6.1.

An example of CMAB: probabilistic maximum coverage (PMC) The input for the PMC problem is a weighted bipartite graph $G = (L, R, E)$, where each edge $(u, v) \in E$ is associated with a weight $\mu(u, v)$. Denote $\mu = (\mu(u, v))_{(u, v) \in E}$ as the edge weight vector. The goal is to find a node set $S \subseteq L$ with $|S| = K$ to maximize the number of influenced nodes in $R$, where each node $v \in R$ can be influenced by $u \in S$ with independent probability $\mu(u, v)$. The advertisement placement problem can be modeled by PMC, where $L$ is the web page set, $R$ is the user set and $\mu(u, v)$ represents the probability that user $v$ clicks the advertisement on web page $u$. In this application, the user click probabilities are unknown and need to be learned during iterative interactions. The PMC problem fits our CMAB framework with each edge being a base arm and edges starting from the same node forming a unit. The expected reward of an action $S$ is the expected number of nodes finally influenced by it, which is defined as

$$r(S, \mu) = \sum_{v \in R} \left( 1 - \prod_{(u, v) \in E, u \in S} (1 - \mu(u, v)) \right). \quad (3)$$

It is proved that the reward function satisfies Assumption 2 [7] and the Greedy oracle can provide an approximate solution with approximation ratio $(1 - \frac{1}{\alpha})$ for any input [23].

4 Algorithm

In this section, we introduce the combinatorial Thompson sampling (CTS) algorithm with Beta priors and Greedy oracle (presented in Algorithm 2) for CMAB problems.

**Algorithm 2 CTS algorithm with Beta priors and Greedy oracle**

1: Input: base arm set $[m]$, unit set $U$, action size $K$
2: Initialize: $\forall i \in [m], a_i = b_i = 1$
3: for $t = 1, 2, \ldots$ do
4: \quad $\forall i \in [m]$: Sample $\theta_{t,i} \sim \text{Beta}(a_i, b_i)$. Denote $\theta_t = (\theta_{t,1}, \theta_{t,2}, \ldots, \theta_{t,m})$
5: Select action $S_t = \text{Greedy}([m], \theta_t, U, K)$ and receive the observation $Q_t$
6: //Update
7: for $(i, X_{t,i}) \in Q_t$ do
8: \quad With probability $X_{t,i}$, $Y_{t,i} = 1$; with probability $1 - X_{t,i}$, $Y_{t,i} = 0$
9: Update $a_i = a_i + Y_{t,i}$; $b_i = b_i + (1 - Y_{t,i})$
10: end for
11: end for

The algorithm maintains a Beta distribution with parameters $a_i$ and $b_i$ for each base arm $i \in [m]$. In the beginning, it initializes $a_i = b_i = 1, \forall i \in [m]$ (Line 2). In each round $t$, the algorithm first samples a parameter candidate $\theta_{t,i}$ from $\text{Beta}(a_i, b_i)$ representing the current estimate for $\mu_i$ (Line 4). Then the Greedy oracle outputs the solution $S_t$ according to the input vector $\theta_t = (\theta_{t,1}, \theta_{t,2}, \ldots, \theta_{t,m})$ (Line 5). Based on the observation feedback, the algorithm then updates the corresponding Beta distributions for observed base arms (Line 7-10).

5
5 Lower Bound

We investigate the hardness of the CTS algorithm to solve CMAB problems with Greedy oracle by proving a problem-dependent regret lower bound.

First, we introduce some notations that will be used in the regret analysis. Recall $S_g$ is the solution returned by the Greedy oracle when the input is $\mu$. We denote it as $S_g = \{s_{g,1}, s_{g,2}, \ldots, s_{g,K}\}$, where $s_{g,k}$ is the $k$-th selected unit by Greedy. Further, define $S_{g,k} = \{s_{g,1}, s_{g,2}, \ldots, s_{g,k}\}$ as the sequence containing the first $k$ units for any $k \in [K]$. Similarly, let $S_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,K}\}$ and $S_{i,k} = \{s_{i,1}, s_{i,2}, \ldots, s_{i,k}\}$. Note $S_{g,0} = S_{i,0} = \emptyset$. The corresponding gaps are defined to measure the hardness of the task and the performance of the algorithm.

**Definition 1.** (Gaps) For any unit $s \in \mathcal{U}$ and index $k \in [K]$ such that $s \notin S_{g,k-1}$, define the marginal reward gap

$$\Delta_{s,k} = r(s_{g,k}, \mu) - r(S_{g,k-1} \cup \{s\}, \mu)$$

as the reward difference between $S_{g,k}$ and $S_{g,k-1} \cup \{s\}$. According to the Greedy algorithm, we have $\Delta_{s,k} > 0$ for any $k$ such that $s \notin S_{g,k}$. And for any action $S \in \mathcal{S}$, define $\Delta_S = \max\{r(S, \mu) - r(S, \mu), 0\}$ as the reward difference from the Greedy’s solution $S_g$. Let

$$\Delta_{s}^{\min} = \min_{S \in \mathcal{S} : s \in S} \Delta_S, \quad \Delta_{s}^{\max} = \max_{S \in \mathcal{S} : s \in S} \Delta_S$$

be the minimum and maximum reward gap of actions containing unit $s$, respectively. Denote $\Delta_{\max}^{\max} = \max_{S \in \mathcal{S}} \Delta_S$ as the maximum reward gap over all suboptimal actions.

We take the following PMC problem (shown in Figure 1) as the instance to carry out the hardness analysis. Each edge in the graph is a base arm and the set of all outgoing edges from a single node forms a unit. The action size is set to $K = 2$. The weight $\mu(u,v)$ of each edge $(u, v)$ is listed on the edges, where we assume $0 < \Delta \leq 0.04$. The expected reward $r(S, \mu)$ of an action $S$ under $\mu$ is defined as Eq (3). For example, when $u_1$ and $u_2$ are selected, the probability of $v_1$ being influenced is $1 - (1 - \mu(u_1,v_1))(1 - \mu(u_2,v_1)) = 0.592$ and the probability that $v_2$ is influenced is $\mu(u_2, v_2) = 0.3$. The expected reward of $S = \{u_1, u_2\}$ is $r(S, \mu) = 0.592 + 0.3 = 0.892$. For simplicity, we also assume the output of each base arm in each round is exactly its mean. This assumption still satisfies the above properties and is also adopted in previous lower bound proofs [3] to simplify the analysis.

![Figure 1: The underlying graph of the PMC instance used to derive the hardness analysis.](image)

For convenience, we first list the expected reward of each action in this problem in Table 1. We can find that the greedy solution is $S_g = \{u_2, u_1\}$ with $s_{g,1} = u_2$, $s_{g,2} = u_1$, and $r(S_g, \mu) = 0.892$, while the optimal action is $\{u_1, u_4\}$. The corresponding marginal reward gaps of each unit can be then computed as follows.

$$\Delta_{u_1,1} = 0.01; \quad \Delta_{u_2,1} = \Delta; \quad \Delta_{u_3,2} = 0.012 + 0.7\Delta;$$

$$\Delta_{u_4,1} = 0.02, \quad \Delta_{u_4,2} = 0.056.$$
Further, the cumulative greedy regret satisfies
\[ R_g(T) = \Omega \left( \frac{\log T}{\Delta^2} \right) = \Omega \left( \frac{\log T}{\Delta^2} \right). \] (5)

The proof of Theorem 1 follows directly the intuition of the above hardness analysis. Due to the space limit, we include the detailed proof in Appendix B. The reason why we consider using Gaussian priors to derive the lower bound analysis is that we hope to use its concentration and anti-concentration inequalities. The analysis can directly apply to other types of prior distributions if similar inequalities can be provided. The main operations of CTS with Gaussian priors are very similar to that of Algorithm 2, while the only difference is on the prior distribution for unknown parameters and the corresponding updating mechanism. To be self-contained, we also present the detailed CTS algorithm with Gaussian priors in Appendix B.

Lin et al. [22] also show a lower bound for greedy regret of order $\Omega(\log T/\Delta^2)$ with $\Delta \in (0, 1/4)$. However, the problem instance used to derive this lower bound is not a CMAB problem and thus their result is not comparable with Theorem 1.

### 6 Upper Bound

By investigating the properties of the CTS algorithm and the Greedy oracle, we also provide a problem-dependent regret upper bound for Algorithm 2 to solve general CMAB problems.

**Theorem 2.** (Upper bound) The cumulative greedy regret of Algorithm 2 can be upper bounded by
\[
R_g(T) \leq \sum_{s \neq s_{g,1}} \max_{k: s \in S_{g,k}} \frac{6B^2 |s|^2 \Delta_{s,k}}{(\Delta_{s,k} - 2B |\cup S_{g}| \varepsilon^2)} \Delta_{\max} + \sum_{k \in [K]} \frac{C^t}{\varepsilon^4} |s_{g,k}| \Delta_{\max}
\]
Thus the \( S \) where the last equality holds since \( \Delta \)

To get an upper bound for the RHS, it requires a sufficient number of selections of the exact optimal action. Thus we can also compare our regret upper bound with the theoretical results of the Greedy oracle in Appendix C, which achieves the same order of the upper bound with Theorem 2 only up to some constant factors.

### 6.1 Discussions

#### Challenges in analyzing the \( \alpha \)-approximate regret with CTS

The \( \alpha \)-approximate regret is first brought up in analyzing UCB-type algorithms [6, 7, 29]. Under UCB, benefiting from the monotonicity between the true parameter \( \mu \) and the UCB parameter \( \tilde{\mu} \), the \( \alpha \)-approximate regret can be deduced as

\[
\alpha \cdot r(S^*, \mu) - r(S_t, \mu) \leq \alpha \cdot r(S^*, \tilde{\mu}) - r(S_t, \tilde{\mu}) \leq r(S_t, \tilde{\mu}) - r(S_t, \mu) \leq \sum_{i \in S_t} |\tilde{\mu}_i - \mu_i|,
\]

where \( S^* \in \arg\max_{S \in S} r(S, \mu) \) is an exact optimal action under real parameter \( \mu \). Thus it only needs to bound the number of selections of bad action \( S_t \) to get an upper bound for the \( \alpha \)-approximate regret. However, under CTS, since there is no monotonicity between the true parameter \( \mu \) and the surrogate parameter \( \theta \), the approximate regret can only be deduced as

\[
\alpha \cdot r(S^*, \mu) - r(S_t, \mu) \leq \alpha \cdot r(S^*, \mu) - \alpha \cdot r(S^*, \theta) + \alpha \cdot r(S^*, \theta) - r(S_t, \mu) \\
\leq \alpha \cdot r(S^*, \mu) - \alpha \cdot r(S^*, \theta) + r(S_t, \theta) - r(S_t, \mu) \\
\leq \alpha \sum_{i \in S^*} |\theta_i - \mu_i| + \sum_{i \in S_t} |\theta_i - \mu_i|.
\]

To get an upper bound for the RHS, it requires a sufficient number of selections of the exact optimal action \( S^* \), which may not be the case with approximate oracles like the example shown in Theorem 1. Thus the \( \alpha \)-approximate regret may not well fit TS-type algorithms.

#### Tightness of the upper bound

We now discuss the tightness of the regret upper bound in Theorem 2 based on the problem instance shown in Figure 1. Specific to this problem, we have \(|s| \leq 2\) for all \( s \) since each node has no more than 2 outgoing edges. And based on [21, Theorem 4], the coefficient of the Lipschitz condition in this problem is \( \beta = 1 \). When \( 0 < \Delta \leq 0.04 \), we have \( \Delta_{s,1} = \min_{k: s \notin S_k} \Delta_{s,k} \) for any \( s \neq s_{g,1} = u_2 \), and \( \Delta_{\max} = \Delta_{\{u_2,u_4\}} = 0.056 \) is a constant. Thus our regret upper bound in this problem instance is of order

\[
O \left( \sum_{s \neq s_{g,1}} \max_{k: s \notin S_k} \frac{B^2 |s|^2 \Delta_{\max} \log T}{\Delta_{s,k}^2} \right) = O \left( \sum_{s \in \{u_1,u_2,u_4\}} \frac{\log T}{\Delta_{s,1}^2} \right) = O \left( \frac{\log T}{\Delta^2} \right),
\]

where the last equality holds since \( \Delta_{u_1,1} = 0.01 \), \( \Delta_{u_4,1} = 0.02 \) are constants and \( \Delta_{u_3,1} = \Delta \).

We can see our regret upper bound matches the lower bound of (5) in Theorem 1 only up to some constant factors in this specific problem instance.

#### Comparison with MAB

When each unit contains only one base arm and the action size is \( K = 1 \), our CMAB framework recovers the MAB problem and the Greedy oracle can provide the exact optimal solution. Thus we can also compare our regret upper bound with the theoretical results of the TS algorithm in MAB problems. In the MAB problem, the expected reward of each action is exactly the mean of the base arm contained in this action. Thus the Lipschitz coefficient is just \( \beta = 1 \) and \(|s| = 1\) for any unit \( s \). The optimal action is \( S_g = S_{g,1} \) with \(|\cup S_g| = 1\). And for any unit \( s \neq s_{g,1} \),
we have $\Delta^\text{max} = \Delta_{s,1}$. Thus, according to (6) of Theorem 2, the regret upper bound of Algorithm 2 in MAB problems is of order $O\left( \sum_{s \neq s_g,1} \log T \right)$, which recovers the main order of the regret upper bound of TS for MAB problems [3].

**Comparison with [22]** Though Lin et al. [22] also study greedy regret, the results are not directly comparable in general since the setting studied in [22] is not a CMAB setting. We find that the PMC problem under a bar graph in these two settings can be equivalent, where a bar graph is a special bipartite graph with each left node’s outdegree being 1 (indegree being 0) and each right node’s indegree being 1 (outdegree being 0). In this case, our greedy regret upper bound is of order $O(m \log T / \Delta^2)$ and theirs is $O(mK \log T / \Delta^2)$. So ours is $O(K)$ better than theirs. Even in this case, their algorithm needs to estimate $O(m \cdot 2^m)$ parameters, while Algorithm 2 is more efficient and only needs to estimate $O(m)$ parameters.

**The definition of the marginal reward gap** Recall that the Greedy oracle provides approximate solutions for problems whose expected reward satisfies monotonicity and submodularity on the action set. Formally, the submodularity means for any action $S \subseteq T$ and unit $s \notin T$, there is $r(S \cup \{s\}, \mu) - r(S, \mu) \geq r(T \cup \{s\}, \mu) - r(T, \mu)$, which characterizes the phenomenon of diminishing returns. One may concern that in these problems, due to the submodularity, the marginal reward gap $\Delta_{s,k}$ for larger $k$ may become much smaller and the main order of the upper bound thus blows up. We clarify that the submodularity cannot imply the relationships among marginal reward gaps $\Delta_{s,k}$ for different $k \in [K]$. The problem instance in Figure 1 satisfying submodularity [16] indicates that the marginal reward gap $\Delta_{s,k}$ does not necessarily decrease with the increase of $k$.

### 6.2 Proof Sketch

In this section, we briefly introduce the proof idea of Theorem 2. At any step $k \in [K]$, for any unit $s \notin S_{g,k}$, the selection of $s$ in action $S_t$ may force the Greedy oracle to choose a worse action in subsequent steps and make CTS suffer constant regret. A sufficient condition for generating zero greedy regret in round $t$ is that each unit $s_{t,k}$ selected in step $k$ is actually $s_{g,k}$ for any $k \in [K]$. Thus to bound the cumulative greedy regret, we sequentially analyze whether $s_{g,k}$ is appropriately selected at each step $k$.

Recall that under the framework of the CTS algorithm, the Greedy oracle sequentially selects $s_{t,k}$ for $k \in [K]$ based on $\theta_t$ sampled from posterior distributions in round $t$. Focusing on step $k = 1$, to ensure $s_{t,1} = s_{g,1}$, the algorithm needs to guarantee the accurate estimations $\theta_{t,s}$ for any unit $s \in U$ such that $r(\{s\}, \theta_t) < r(S_{g,1}, \theta_t)$, $s \neq s_{g,1}$, where $\theta_{t,s}$ is the projection of $\theta_t$ on unit $s$. We first assume $s_{g,1}$ is already estimated well. Then based on this assumption, when all of the other units $s \neq s_{g,1}$ have been explored $O(\log T / \Delta_{g,1}^2)$ times and thus estimated accurately, $s_{g,1}$ would be selected at the first step with high probability. But if after any other unit $s \neq s_{g,1}$ has already been estimated well, $s_{g,1}$ is still not selected appropriately, we can conclude that the estimations for $s_{g,1}$ are not accurate enough. In this case, the Beta posterior for $s_{g,1}$ tends to be uniformly distributed. When CTS sample $\theta_{t,s_{g,1}}$ from its Beta posterior, with constant probability there would be $r(S_{g,1}, \theta_t) > r(\{s\}, \theta_t)$ for any unit $s \neq s_{g,1}$. Thus after some rounds, $s_{g,1}$ would be selected for enough times and also estimated accurately. In the following rounds, $s_{g,1}$ would be selected appropriately at the first step with high probability. Above all, the expected number of misselections at the first step can be bounded.

The above analysis can apply to other cases when $k = 2, 3, \ldots, K$. Based on the correct selections in the first $k - 1$ steps, the misselection of $s_{t,k}$ also comes from the bad estimations for both $s_{g,k}$ and other units $s \notin S_{g,k}$. To distinguish $r(S_{g,k-1} \cup \{s\}, \theta_t)$ from $r(S_{g,k}, \theta_t)$, those units need to be explored at least $O(\log T / \Delta_{s,k}^2)$ times.

According to the above analysis, for each unit $s \neq s_{g,1}$, we define the exploration price as

$$L(s) = O\left( \max_{k: s \notin S_{g,k}} \frac{\log T}{\Delta_{s,k}^2} \right).$$

(8)

To avoid being incorrectly selected at some step $k$, each unit $s \notin S_{g,k}$ needs to be explored for at least $L(s)$ times. The sum of $L(s)$ over all units $s \neq s_{g,1}$ leads to the main order of the regret upper bound in Theorem 2.
6.3 Extension to Multiple-solution Case

We can also extend the analysis of the regret upper bound to the case where multiple solutions may be returned by \texttt{Greedy} with input \( \mu \), or equivalently the optimal unit in each step \( k \) (Line 4 in Algorithm 1) may not be unique. Let

\[
\sigma_K = \left\{ \{s_1, s_2, \ldots, s_K\} : s_1 \in \arg\max_s r(\{s\}, \mu), \ldots, s_K \in \arg\max_{s \not\in \{s_1, \ldots, s_{K-1}\}} r(\{s_1, \ldots, s\}, \mu) \right\}
\]

be the set of all actions that are possibly returned by the \texttt{Greedy} oracle when the input is \( \mu \). Here we do not care how \texttt{Greedy} breaks the tie at each step and regard this process as a black box. In order to take into account the worst case where \texttt{Greedy} always returns a solution with minimum reward compared to other possible solutions, we define \( S_g \in \arg\min_{S \in \sigma_K} r(S, \mu) \) as one of \texttt{Greedy}'s possible solutions with the minimum expected reward and consider the cumulative greedy regret defined in Eq (2).

The regret analysis of Algorithm 2 in this case is similar to the proof of Theorem 2. A sufficient condition for generating zero regret in round \( t \) is that the selected action \( S_t \) falls into the set \( \sigma_K \). Thus to bound the cumulative greedy regret, we sequentially analyze whether the unit \( s_{t,k} \) selected in each step \( k \) is an optimal unit conditioned on the previously selected units \( S_{t,k-1} \). For completeness, we include the regret upper bound as well as the detailed proof for this case in Appendix E.

7 Conclusion

In this paper, we aim to answer the question of whether the convergence analysis of TS can be extended beyond exact oracles in the CMAB area. Taking the common offline (approximation) \texttt{Greedy} oracle as an example, we derive the hardness analysis of CTS for CMAB problems based on a constructed CMAB problem instance. When using CTS with \texttt{Greedy} oracle to solve this problem, we find that the algorithm needs to explore at least \( \Omega(\log T/\Delta^2) \) rounds to distinguish suboptimal units from the optimal unit at some step. However, at least constant regret needs to be paid for each exploration round. The mismatch between the gap to be distinguished and the actually paid regret forces the algorithm to pay the cumulative greedy regret of order \( \Omega(\log T/\Delta^2) \). We also provide an almost matching problem-dependent regret upper bound for the CTS algorithm with \texttt{Greedy} oracle to solve CMAB problems. The upper bound is tight on the constructed problem instance only up to some constant factors and also recovers the main order of TS when solving MAB problems.

An interesting future direction is to extend the current CMAB framework to the case with probabilistically triggered arms (CMAB-T). The CMAB-T framework can model the OIM problem on general social networks. As shown in [21], using UCB-type algorithms to solve such a problem may face great challenges on the computation efficiency. This problem is expected to be avoided by TS-type algorithms since TS would sample candidate parameters to escape the computation of complicated optimization problems. However, the current proof idea based on each selection step of the \texttt{Greedy} oracle (proof of Lemma 1) cannot directly apply to this setting as different units may probabilistically trigger some common base arms. New proof techniques are required to derive the theoretical guarantee of the TS algorithm with the \texttt{Greedy} oracle in this framework.
Acknowledgement

The corresponding author Shuai Li is supported by National Natural Science Foundation of China (62006151, 62076161). This work is sponsored by Shanghai Sailing Program.

References


A Notations

Before the proofs, we first introduce some notations that would be used in the regret analysis.

For any base arm \( i \in [m] \), let \( N_{t,i} = \sum_{\tau < t} 1 \{ i \in \cup S_{\tau} \} \) be the number of observations of \( i \) and 
\[ \hat{\mu}_{t,i} = \frac{1}{N_{t,i}} \sum_{\tau < t; i \in S_{\tau}} X_{\tau,i} \]
be the empirical mean outcome of \( i \) before the start of round \( t \). Denote 
\( \hat{\mu} = (\hat{\mu}_{t,1}, \hat{\mu}_{t,2}, \ldots, \hat{\mu}_{t,m}) \) as the empirical mean vector. Since a unit of base arms will be selected together, we abuse a bit the notation \( N_{t,s} = \sum_{\tau < t} 1 \{ s \in S_{\tau} \} \) to represent the number of selections of unit \( s \) before the start of round \( t \). For any mean vector \( \theta \in [0,1]^m \), denote \( \theta_s \) and \( \theta_S \) as the projection of \( \theta \) on unit \( s \) and action \( S \).

B Proof of Theorem 1

To be self-contained, we present the detailed CTS algorithm with Gaussian priors and Greedy oracle in Algorithm 3. The main operation of the algorithm is the same as Algorithm 2, the only difference is on the prior distribution for unknown parameters and the corresponding updating mechanism. As we previously discussed, the reason for using the Gaussian distribution to derive the lower bound analysis is that we hope to use its concentration and anti-concentration inequalities. The analysis directly applies to other types of prior distributions if similar inequalities can be provided.

In this algorithm, an initialization phase is introduced to ensure at least one observation has been collected for each base arm (Line 2). In each round \( t \), the posterior for \( \mu_i \) is given by \( N(\hat{\mu}_{t,i}, \frac{1}{N_{t,i}}) \), a Gaussian distribution with mean \( \hat{\mu}_{t,i} \) and variance \( \frac{1}{N_{t,i}} \). If arm \( i \) is observed in this round, the mean and variance of its Gaussian distribution will be updated as Line 7.

**Algorithm 3 CTS algorithm with Gaussian priors and Greedy oracle**

1: Input: base arm set \([m]\), unit set \(U\), action size \(K\)
2: Initialization: for each unit \( s \in U \), select an arbitrary action \( S \in S \) such that \( s \in S \). Update \( N_{1,s} \) and \( \hat{\mu}_{1,i} \) for any \( i \in [m] \) based on observations in this phase.
3: for \( t = 1, 2, \ldots \) do
4:   \( \forall i \in [m] \): Sample \( \theta_{t,i} \sim N(\hat{\mu}_{t,i}, \frac{1}{N_{t,i}}) \). Denote \( \theta_t = (\theta_{t,1}, \theta_{t,2}, \ldots, \theta_{t,m}) \)
5:   Select action \( S_t = \text{Greedy}([m], \theta_t, U, K) \) and receive the observation \( Q_t \)
6:   for \( (i, X_{t,i}) \in Q_t \) do
7:     \( \hat{\mu}_{t+1,i} = \frac{\hat{\mu}_{t,i} N_{t,i} + X_{t,i}}{N_{t+1,i} + 1}, N_{t+1,i} = N_{t,i} + 1 \)
8:   end for
9: end for

For any unit \( s \neq s_{g,1} \), in any round \( t \), define event
\[
A_s(t) = \left\{ N_{t,s} \geq 2\frac{|s|}{7\Delta_{s,1}^2} \log T + \log \log T \right\}
\]
When \( \mathbb{P}(A_s(t)) \geq \frac{1}{2} \), we have
\[
\mathbb{E}[N_s(T + 1)] \geq \mathbb{E}[N_s(t)] \geq \mathbb{E}[N_s(t) \mid A_s(t)] \cdot \mathbb{P}(A_s(t))
\]
\[
\geq \frac{1}{2} \cdot 2\frac{|s|}{7\Delta_{s,1}^2} \log T + \log \log T
\]
\[
= \Omega \left( \frac{\log T}{\Delta_{s,1}^2} \right),
\]
when \( T \) is sufficiently large.

If \( \mathbb{P}(A_s(t)) < \frac{1}{2} \) in round \( t \), we then prove unit \( s \) will be selected by Greedy in the following rounds with a large probability and the expected number of its selections will finally exceed this threshold.
\[
\mathbb{P}(s \in S_t) \geq \mathbb{P}(s = s_{t,1})
\]
where the last equality holds since different units have no common base arms and thus the events on
their rewards are independent of that on others.

For the first part, we have
\[
\begin{align*}
\mathbb{P}(r(\{s\}, \theta_t) > r(S_{g,1}, \mu)) &= \mathbb{P}(r(\{s\}, \theta_t) - r(\{s\}, \mu) > r(S_{g,1}, \mu) - r(\{s\}, \mu)) \\
&= \mathbb{P}\left( \sum_{i \in s} (\theta_{t,i} - \mu_i) > \Delta_{s,1} \right) \\
&\geq \mathbb{P}\left( \sum_{i \in s} (\theta_{t,i} - \mu_i) > \Delta_{s,1} \right) \mathbb{P}(\neg A_s(t)) \\
&\geq \frac{1}{2} \mathbb{P}\left( \sum_{i \in s} (\theta_{t,i} - \mu_i) \sqrt{\frac{N_{t,s}}{|s|}} > \Delta_{s,1} \sqrt{\frac{N_{t,s}}{|s|}} \right) \\
&> \frac{1}{16\sqrt{\pi}} \exp \left( -\frac{7\Delta_{s,1}^2}{2|s|} \right) \times \left( \frac{2|s|}{7\Delta_{s,1}^2} \log T + \log \Delta_{s,1}^2 - \log \log T \right) \\
&= \frac{1}{16\sqrt{\pi}} \frac{\log T}{T \cdot \Delta_{s,1}^2},
\end{align*}
\]

where the third equality is due to the reward definition in this specific problem instance and the last
inequality is due to the Lemma 6 and the event $\neg A_s(t)$.

For the second part, we have
\[
\begin{align*}
\mathbb{P}(\forall s' \neq s : r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)) &= \mathbb{P}(r(S_{g,1}, \theta_t) \leq r(S_{g,1}, \mu)) \cdot \mathbb{P}(\forall s' \neq \{s, s_{g,1}\} : r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)) \\
&= \mathbb{P}\left( \sum_{i \in s_{g,1}} \theta_{s,i} \leq \sum_{i \in s_{g,1}} \mu_i \right) \cdot \mathbb{P}(\forall s' \neq \{s, s_{g,1}\} : r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)) \\
&\geq \mathbb{P}(\forall i \in s_{g,1} : \theta_{t,i} \leq \mu_i) \cdot \mathbb{P}(\forall s' \neq \{s, s_{g,1}\} : r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)) \\
&\geq \left( \frac{1}{2} \right)^{|s|_{g,1}} \mathbb{P}(\forall s' \neq \{s, s_{g,1}\} : r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)) \\
&= \left( \frac{1}{2} \right)^{|s|_{g,1}} \prod_{s' \neq \{s, s_{g,1}\}} \mathbb{P}(r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)),
\end{align*}
\]

where the first and last equality is again due to the independence over different units, the last inequality
comes from the result of Lemma 6 and the independence over base arms. For each term in the last
formula, there is
\[
\begin{align*}
\mathbb{P}(r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)) &= \mathbb{P}(r(\{s\}', \theta_t) \leq r(S_{g,1}, \mu)) \\
&= \mathbb{P}\left( \sum_{i \in s'} (\theta_{t,i} - \mu_i) \leq \Delta_{s',1} \right) \\
&= \mathbb{P}\left( \sum_{i \in s'} (\theta_{t,i} - \mu_i) \right) \leq \Delta_{s',1} \sqrt{\frac{N_{t,s'}}{|s'|}} \\
&= \mathbb{P}\left( \sum_{i \in s'} (\theta_{t,i} - \mu_i) \right) \leq \Delta_{s',1} \sqrt{\frac{N_{t,s'}}{|s'|}}
\end{align*}
\]
\[ \geq 1 - \frac{1}{4} \exp \left( - \frac{\Delta^2_{s,1} N_{t,s'}}{2|s'|} \right) \]
\[ \geq 1 - \frac{1}{4} \exp \left( - \frac{\Delta^2_{s,1}}{2|s'|} \right), \]

where the first inequality comes from Lemma 6 and the last one is because \( N_{t,s} \geq 1 \) after the initialization phase for any \( s \). Thus we have

\[ \mathbb{P}(\forall s' \neq s : r(\{s'\}, \theta(t)) \leq r(S_{g,1}, \mu)) \geq \left( \frac{1}{2} \right)^{|s_{g,1}|} \prod_{s' \notin \{s, s_{g,1}\}} \left( 1 - \frac{1}{4} \exp \left( - \frac{\Delta^2_{s,1}}{2|s'|} \right) \right) \triangleq C'', \]

here \( C'' \) can be regarded as a constant. Above all, we have

\[ \mathbb{P}(s \in S_t) \geq \frac{1}{16\sqrt{T} \pi} \log T \cdot \frac{\log T}{\Delta_{s,1}^2} \triangleq p, \]

the total expected number of rounds when unit \( s \) is selected in \( S_t \) is at least

\[ \mathbb{E}[N_{T+1,s}] \geq Tp = \Omega \left( \frac{\log T}{\Delta_{s,1}^2} \right). \]

Thus we have proved (4) in Theorem 1.

And in the above CMAB instance, we can find that

\[ \Delta_{u_3}^{\min} = 0.52\Delta + 0.008 = 0.52\Delta_{u_3,1} + 0.008, \]

Above all,

\[ R_g(T) = \mathbb{E} \left[ \sum_{t=1}^{T} \Delta_{S_t} \right] \geq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}\{u_3 \in S_t\} \Delta_{S_t} \right] \]
\[ \geq \mathbb{E} [N_{T+1,u_3}] \Delta_{u_3,\min} \]
\[ = \Omega \left( \frac{\log T}{\Delta_{u_3,1}^2} \cdot (0.52\Delta_{u_3,1} + 0.008) \right) \]
\[ = \Omega \left( \frac{\log T}{\Delta_{u_3,1}^2} \right). \]

We then complete the proof of (5) in Theorem 1.

**C Proof of Theorem 2**

For any unit \( s \neq s_{g,1} \), we define the exploration price \( L(s) \) as

\[ L(s) = \max_{k: s \notin S_{s,k}} \frac{6B^2 |s|^2 \log T}{(\Delta_{s,k} - 2B |\cup S_g| \varepsilon)^2}. \]

In each round \( t \), define the following event

\[ B(t) = \left\{ \exists i \in [m] : |\theta_{t,i} - \mu_{t,i}| > \sqrt{\frac{3 \log T}{2N_{t,i}}} \right\}, \]
\[ C(t) = \left\{ \exists i \in [m] : |\hat{\mu}_{t,i} - \mu_i| > \sqrt{\frac{3 \log T}{2N_{t,i}}} \right\}. \]

The greedy regret then can be decomposed as

\[ R_g(T) \leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}\{\neg B(t), \neg C(t)\} \Delta_{S_t} \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}\{B(t)\} \Delta_{S_t} \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}\{C(t)\} \Delta_{S_t} \right]. \]

(9)
We then bound these three terms in (9) one by one.

The first term in (9):
Recall $\Delta_{S_g} = \max \{r(S_g, \mu) - r(S_t, \mu), 0\}$. If $s_{t,k} = s_{g,k}$ for any $k \in [K]$, we must have $\Delta_{S_g} = 0$. Thus to bound this term, we analyze the relationship between $s_{t,k}$ and $s_{g,k}$ sequentially for $k = 1, 2, \ldots, K$. According to this idea, this term can be bounded by

$$E \left[ \sum_{t=1}^{T} \mathbb{I} \{ -B(t), -C(t) \} \Delta_{S_t} \right]$$

$$\leq \sum_{k \in [K]} E \left[ \sum_{t=1}^{T} \mathbb{I} \{ -B(t), -C(t), s_{t,k-1} = s_{g,k-1}, \| \theta_{t,s_{g,k-1}}, \mu_{s_{g,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \neq s_{g,k} \} \Delta_{S_t} \right]$$

$$+ \sum_{k \in [K]} E \left[ \sum_{t=1}^{T} \mathbb{I} \{ s_{t,k} = s_{g,k}, \| \theta_{t,s_{g,k}}, \mu_{s_{g,k}} \|_\infty > \varepsilon \} \right] \Delta_{\max}.$$

According to Lemma 1, the first term can be bounded by

$$\sum_{k \in [K]} E \left[ \mathbb{I} \{ -B(t), -C(t), s_{t,k-1} = s_{g,k-1}, \| \theta_{t,s_{g,k-1}}, \mu_{s_{g,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \neq s_{g,k} \} \Delta_{S_t} \right]$$

$$\leq \sum_{k \in [K]} \mathbb{E} \left[ \sum_{s \not\in S_{g,k}} \sum_{t=1}^{T} \mathbb{I} \{ s = s_{t,k}, N_{t,s} \leq L(s) \} \Delta_{S_t} \right] + \sum_{k \in [K]} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{g,k}|} \Delta_{\max}$$

$$\leq \mathbb{E} \left[ \sum_{s \not\in S_{g,k}} \sum_{t=1}^{T} \mathbb{I} \{ s = s_{t,k}, N_{t,s} \leq L(s) \} \Delta_{S_t} \right] + \sum_{k \in [K]} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{g,k}|} \Delta_{\max}$$

$$\leq \sum_{s \not\in S_{g,k}} \frac{L(s)}{\Delta_{\max}} + \sum_{k \in [K]} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{g,k}|} \Delta_{\max}$$

$$\leq \sum_{s \not\in S_{g,k}} \max_{k \in [K]} \frac{6B^2|s|^2 \Delta_{\max} \log T}{(\Delta_{s,k} - 2B|\cup S_g|\varepsilon)^2} + \sum_{k \in [K]} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{g,k}|} \Delta_{\max},$$

where $C, C'$ are two universal constants. According to Lemma 3, the second term can be bounded by

$$\sum_{k \in [K]} \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I} \{ s_{t,k} = s_{g,k}, \| \theta_{t,s_{g,k}}, \mu_{s_{g,k}} \|_\infty > \varepsilon \} \right] \Delta_{\max} \leq \sum_{k \in [K]} |s_{g,k}| \left( 2 + \frac{8}{\varepsilon^2} \right) \Delta_{\max}$$

$$= |\cup S_g| \left( 2 + \frac{8}{\varepsilon^2} \right) \Delta_{\max}$$

Above all, for the first term in (9), we have

$$E \left[ \sum_{t=1}^{T} \mathbb{I} \{ -B(t), -C(t) \} \Delta_{S_t} \right]$$

$$\leq \sum_{s \not\in S_{g,k}} \max_{k \in [K]} \frac{6B^2|s|^2 \Delta_{\max} \log T}{(\Delta_{s,k} - 2B|\cup S_g|\varepsilon)^2} + \left( \sum_{k \in [K]} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{g,k}|} + |\cup S_g| \left( 2 + \frac{8}{\varepsilon^2} \right) \right) \Delta_{\max}.$$
where (11) is obtained by Lemma 5.

The third term in (9):

\[
\begin{align*}
&\leq \mathbb{E}\left[ \sum_{t=1}^{T} \mathbb{I}\left\{ \exists i \in [m] : |\theta_{t,i} - \hat{\mu}_{t,i}| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \right\} \right] \cdot \Delta_{\text{max}} \\
&\leq \sum_{i \in [m]} \mathbb{E}\left[ \sum_{t=1}^{T} \mathbb{I}\left\{ |\theta_{t,i} - \hat{\mu}_{t,i}| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \right\} \right] \cdot \Delta_{\text{max}} \\
&= \sum_{i \in [m]} \sum_{t=1}^{T} \sum_{w=1}^{T-1} \mathbb{P}\left( N_{t,i} = w, |\theta_{t,i} - \hat{\mu}_{t,i}| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \right) \cdot \Delta_{\text{max}} \\
&= \sum_{i \in [m]} \sum_{t=1}^{T} \sum_{w=1}^{T-1} \mathbb{P}(N_{t,i} = w) \cdot \mathbb{P}\left( |\theta_{t,i} - \hat{\mu}_{t,i}| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \mid N_{t,i} = w \right) \cdot \Delta_{\text{max}} \\
&\leq \sum_{i \in [m]} \sum_{t=1}^{T} \sum_{w=1}^{T-1} \mathbb{P}(N_{t,i} = w) \cdot 2 \exp (-3 \log T) \cdot \Delta_{\text{max}} \\
&\leq \sum_{i \in [m]} \sum_{t=1}^{T} \frac{2}{T} \cdot \Delta_{\text{max}} \\
&= 2m \Delta_{\text{max}},
\end{align*}
\]

where (10) comes from the result of Lemma 7.

The third term in (9):

\[
\begin{align*}
&\leq \mathbb{E}\left[ \sum_{t=1}^{T} \mathbb{I}\left\{ \exists i \in [m] : |\hat{\mu}_{t,i} - \mu_i| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \right\} \right] \cdot \Delta_{\text{max}} \\
&\leq \sum_{i \in [m]} \mathbb{E}\left[ \sum_{t=1}^{T} \mathbb{I}\left\{ |\hat{\mu}_{t,i} - \mu_i| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \right\} \right] \cdot \Delta_{\text{max}} \\
&= \sum_{i \in [m]} \sum_{t=1}^{T} \sum_{w=1}^{T-1} \mathbb{P}\left( N_{t,i} = w, |\hat{\mu}_{t,i} - \mu_i| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \right) \cdot \Delta_{\text{max}} \\
&= \sum_{i \in [m]} \sum_{t=1}^{T} \sum_{w=1}^{T-1} \mathbb{P}(N_{t,i} = w) \cdot \mathbb{P}\left( |\hat{\mu}_{t,i} - \mu_i| > \sqrt{\frac{3 \log T}{2 N_{t,i}}} \mid N_{t,i} = w \right) \cdot \Delta_{\text{max}} \\
&\leq \sum_{i \in [m]} \sum_{t=1}^{T} \sum_{w=1}^{T-1} \mathbb{P}(N_{t,i} = w) \cdot 2 \exp (-3 \log T) \cdot \Delta_{\text{max}} \\
&\leq \sum_{i \in [m]} \sum_{t=1}^{T} \frac{2}{T} \cdot \Delta_{\text{max}} \\
&= 2m \Delta_{\text{max}},
\end{align*}
\]

where (11) is obtained by Lemma 5.

Combine these three terms in (9), we can get the result in Theorem 2.

\[
R_g(T) \leq \sum_{s \neq s_{\text{c},k}} \max_{\varphi \in \Phi} \frac{6B^2 |s|}{(\Delta_{s,k} \max) T} \mathbb{E}\left[ \sum_{k \in [K]} \left( \frac{C}{\varepsilon^2} \right)^{|s_{\varphi,k}|} \right] \cdot \Delta_{\text{max}} \\
+ \left( \left| S_g \right| \left( \frac{2}{\varepsilon^2} + \frac{8}{\varepsilon^2} \right) + 4m \right) \Delta_{\text{max}},
\]

17
C.1 Technical Lemmas

Lemma 1. In Algorithm 2, for any $k \in [K]$, we have

\[
E \left[ \sum_{t=1}^{T} I\{\neg B(t), \neg C(t), S_{t,k-1} = S_{g,k-1}, \|\theta_{t,S_{g,k-1}} - \mu S_{g,k-1}\|_{\infty} \leq \varepsilon, s_{t,k} \neq s_{g,k} \} \Delta_{s_{t,k-1}} \right] 
\]

\[
\leq E \left[ \sum_{s \notin S_{g,k}} \sum_{t=1}^{T} I\{s = s_{t,k}, N_{t,s} \leq L(s)\} \Delta_{s_{t,k-1}} \right] + C \left( \frac{C'}{\varepsilon^2} \right) |s_{g,k}| \Delta_{\max},
\]

where $C, C'$ are two universal constants.

Proof. Recall for any unit $s \notin S_{g,k}$, $\Delta_{s,k} = r(S_{g,k}, \mu) - r(S_{g,k-1} \cup \{s\}, \mu)$. Define the event

\[
D_k(t) = \left\{ B \sum_{i \in s_{t,k}} |\theta_{t,i} - \mu_i| > \Delta_{s_{t,k-1}} - B \left( 2 \sum_{k' < k} |s_{g,k'}| + |s_{g,k}| + 1 \right) \varepsilon \right\}.
\]

Then the formula in Lemma 1 can further bounded by

\[
E \left[ \sum_{t=1}^{T} I\{\neg B(t), \neg C(t), S_{t,k-1} = S_{g,k-1}, \|\theta_{t,S_{g,k-1}} - \mu S_{g,k-1}\|_{\infty} \leq \varepsilon, s_{t,k} \neq s_{g,k} \} \Delta_{s_{t,k-1}} \right] 
\]

\[
\leq E \left[ \sum_{t=1}^{T} I\{\neg B(t), \neg C(t), S_{t,k-1} = S_{g,k-1}, \|\theta_{t,S_{g,k-1}} - \mu S_{g,k-1}\|_{\infty} \leq \varepsilon, s_{t,k} \neq s_{g,k}, D_k(t) \} \Delta_{s_{t,k-1}} \right] 
\]

\[
+ E \left[ \sum_{t=1}^{T} I\{S_{t,k-1} = S_{g,k-1}, \|\theta_{t,S_{g,k-1}} - \mu S_{g,k-1}\|_{\infty} \leq \varepsilon, s_{t,k} \neq s_{g,k}, \neg D_k(t) \} \right] \Delta_{\max}.
\]

(12)

(13)

For term (12), we claim that the event $\{\neg B(t), \neg C(t), S_{t,k-1} = S_{g,k-1}, \|\theta_{t,S_{g,k-1}} - \mu S_{g,k-1}\|_{\infty} \leq \varepsilon, s_{t,k} \neq s_{g,k}, D_k(t)\}$ implies $N_{t,s_{t,k}} \leq L(s_{t,k})$. This claim can be proved by contradiction.

Suppose $N_{t,s_{t,k}} > L(s_{t,k})$, then we must have

\[
B \sum_{i \in s_{t,k}} |\theta_{t,i} - \mu_i| \leq B \sum_{i \in s_{t,k}} \sqrt{\frac{6 \log T}{N_{t,s_{t,k}}}} 
\]

\[
< B |s_{t,k}| \sqrt{\frac{6 \log T}{6B^2 |s_{t,k}|^2 \log T}} \left( \Delta_{s_{t,k-1}} - B |\cup S_g| \varepsilon \right) 
\]

\[
\leq B |s_{t,k}| \sqrt{\frac{\log T}{B^2 |s_{t,k}|^2 \log T}} \left( \Delta_{s_{t,k-1}} - B \left( 2 \sum_{k' < k} |s_{g,k'}| + |s_{g,k}| + 1 \right) \varepsilon \right) 
\]

\[
= \Delta_{s_{t,k-1}} - B \left( 2 \sum_{k' < k} |s_{g,k'}| + |s_{g,k}| + 1 \right) \varepsilon,
\]

where the first inequality is due to the event of $\neg B(t)$ and $\neg C(t)$, the second inequality comes from the fact $N_{t,s_{t,k}} > L(s_{t,k})$ and the definition of $L(s_{t,k})$. Thus we conclude the event $D_k(t)$ will not happen and the claim is proved.

Then according to the above claim, there is

\[
(12) \leq E \left[ \sum_{t=1}^{T} I\{N_{t,s_{t,k}} \leq L(s_{t,k}), s_{t,k} \notin S_{g,k} \} \Delta_{s_{t,k-1}} \right] 
\]

\[
\leq E \left[ \sum_{s \notin S_{g,k}} \sum_{t=1}^{T} I\{s = s_{t,k}, N_{t,s} \leq L(s)\} \Delta_{s_{t,k-1}} \right].
\]
For term (13), we first define event $E_{k,1}(t)$ as

$$E_{k,1}(t) = \{ \forall \theta' \text{ with } \theta'_i = \theta_{t,i} \text{ for any } i \notin s_{g,k} \text{ and } \|\theta'_{s_{g,k}} - \mu_{s_{g,k}}\|_\infty \leq \varepsilon, \text{ then } s_{g,k} \text{ is the } k\text{-th selected unit by Greedy when the input is } \theta' \}$$

and the event $E_{k,2}(t)$ as

$$E_{k,2}(t) = \{ \|\theta_{t,s_{g,k}} - \mu_{s_{g,k}}\|_\infty > \varepsilon \}.$$

We claim that if the event $\{ S_{t,k-1} = S_{g,k-1}, \|\theta_{t,s_{g,k-1}} - \mu_{s_{g,k-1}}\|_\infty \leq \varepsilon, s_{t,k} \neq s_{g,k}, -D_k(t) \}$ happens, then $E_{k,1}(t)$ and $E_{k,2}(t)$ hold.

We first consider event $E_{k,1}(t)$. To show this event holds, it is sufficient to prove for any $\theta'$ satisfies the condition defined in $E_{k,1}(t)$, $S_{g,k-1}$ is still the set of units selected by Greedy in the first $k-1$ steps under $\theta'$ and for any $s' \notin S_{g,k}$, $r(S_{g,k-1} \cup \{ s' \}, \theta') < r(S_{g,k}, \theta')$.

We now prove that for any $\theta'$ satisfying the condition defined in $E_{k,1}(t)$, $S_{g,k-1}$ is still the set of units selected by Greedy under $\theta'$ in the first $k-1$ steps. The event $S_{t,k-1} = S_{g,k-1}$ means that for any $k' < k, s \notin S_{g,k'}$, we have $r(S_{g,k'}, \theta_i) > r(S_{g,k-1} \cup \{ s \}, \theta_i)$. The mean vector $\theta'$ and $\theta_i$ are only different on $s_{g,k}$, thus for any $k' < k, s \notin S_{g,k'} \cup \{ s_{g,k} \}$, we still have $r(S_{g,k'}, \theta') > r(S_{g,k-1} \cup \{ s \}, \theta')$. As for the unit $s_{g,k}$, for any $k' < k$, we have

$$r(S_{g,k'}, \theta') = r(S_{g,k'}, \theta_i)$$

The last inequality holds due to the requirement of $\varepsilon$ in Theorem 2. Above all, we conclude $\forall k' < k, s \notin S_{g,k'}, r(S_{g,k'}, \theta') > r(S_{g,k-1} \cup \{ s \}, \theta')$, thus $S_{g,k-1}$ is still the set of units selected by Greedy in the first $k-1$ steps.

Next we prove for any $\theta'$ defined in $E_{k,1}(t)$ and unit $s' \notin S_{g,k}$, $r(S_{g,k-1} \cup \{ s' \}, \theta') < r(S_{g,k}, \theta')$.

$$r(S_{g,k-1} \cup \{ s' \}, \theta') = r(S_{g,k-1} \cup \{ s' \}, \theta_i)$$

The first equality is because $\theta'_i = \theta_{t,i}$ for any $i \in s'$ and $i \in S_{g,k-1}$, the third inequality is because $S_{t,k-1} = S_{g,k-1}$ and $\|\theta_{t,s_{g,k-1}} - \mu_{s_{g,k-1}}\|_\infty \leq \varepsilon$. (14) comes from the definition of $-D_k(t)$ and (15) is due to the definition of $\Delta_{s_{g,k}}$ and the fact $S_{t,k-1} = S_{g,k-1}$.
Above all, we have proved if event \( \{S_{t,k-1} = S_{g,k-1}, \|\theta_t, S_{g,k-1} - \mu S_{g,k-1}\|_\infty \leq \varepsilon, s_{t,k} \neq s_{g,k}, -D_k(t) \} \) happens, then \( E_{k,1}(t) \) holds.

Next we consider \( E_{k,2}(t) \). By contradiction, when the event \( \{S_{t,k-1} = S_{g,k-1}, \|\theta_t, S_{g,k-1} - \mu S_{g,k-1}\|_\infty \leq \varepsilon, s_{t,k} \neq s_{g,k}, -D_k(t) \} \) happens, if \( -E_{k,2}(t) = \{\|\theta_{t,q,k} - \mu_{t,q,k}\|_\infty \leq \varepsilon\} \) holds, then \( \theta_t \) satisfies the property of \( \theta' \) defined in \( E_{k,1}(t) \). Thus according to \( E_{k,1}(t) \), \( s_{g,k} \) would be the \( k \)-th selected unit by Greedy when the input is \( \theta_t \), or in other words \( s_{g,k} = s_{t,k} \). This contradicts \( \{S_{t,k-1} = S_{g,k-1}, \|\theta_t, S_{g,k-1} - \mu S_{g,k-1}\|_\infty \leq \varepsilon, s_{t,k} \neq s_{g,k}, -D_k(t) \} \). Thus we conclude \( E_{k,2}(t) \) also holds.

Above all, for term (13) we have

\[
(13) \leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}_{\{E_{k,1}(t), E_{k,2}(t)\}} \right] \Delta_{\max} \\
\leq \sum_{q \geq 0} \mathbb{E} \left[ \sum_{t=\tau_{k,q}+1}^{\tau_{k,q}+1} \mathbb{1}_{\{E_{k,1}(t), E_{k,2}(t)\}} \right] \Delta_{\max} \\
= \left( \sum_{q \geq 0} \mathbb{E} \left[ \sup_{t \geq \tau_{k,q}+1} \prod_{i \in s_{g,k}} \mathbb{P}(\|\theta_{t,i} - \mu_i\| \leq \varepsilon \mid \mathcal{H}_t) \right] - 1 \right) \Delta_{\max} \\
\leq \left( \sum_{q=0}^{[8/\varepsilon^2]-1} (c\varepsilon^{-4})^{s_{g,k}} + \sum_{q \geq [8/\varepsilon^2]} e^{-q^2/8} (c'\varepsilon^{-4})^{s_{g,k}} \right) \Delta_{\max} \\
\leq \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{s_{g,k}} \Delta_{\max},
\]

where \( \tau_{k,q} \) is the round at which \( E_{k,1}(t) \cap -E_{k,2}(t) \) occurs for the \( q \)-th time, note \( \tau_{k,0} = 0 \) for any \( k \in [K] \). The third inequality comes from the result in Lemma 2 and the fourth comes from the Lemma 8. Here \( C, C' \) are two universal constants. \( \square \)

**Lemma 2.** Let \( \tau_{k,q} \) be the round at which \( E_{k,1}(t) \cap -E_{k,2}(t) \) occurs for the \( q \)-th time, let \( \tau_{k,0} = 0 \) for any \( k \in [K] \). Then for Algorithm 2, we have

\[
\mathbb{E} \left[ \sum_{t=\tau_{k,q}+1}^{\tau_{k,q}+1} \mathbb{1}_{\{E_{k,1}(t), E_{k,2}(t)\}} \right] \leq \mathbb{E} \left[ \sup_{t \geq \tau_{k,q}+1} \prod_{i \in s_{g,k}} \mathbb{P}(\|\theta_{t,i} - \mu_i\| \leq \varepsilon \mid \mathcal{H}_t) \right] - 1.
\]

**Proof.** Conditioned on history \( \mathcal{H}_t \), the event \( E_{k,1}(t) \) and \( E_{k,2}(t) \) are independent. Thus

\[
\mathbb{E} \left[ \sum_{t=\tau_{k,q}+1}^{\tau_{k,q}+1} \mathbb{1}_{\{E_{k,1}(t), E_{k,2}(t)\}} \right] = \mathbb{E} \left[ \sum_{q \geq 1} (q' - 1) \mathbb{P}(E_{k,2}(\tau_{k,q},q') \mid \mathcal{H}_{\tau_{k,q},q'}) \prod_{j=1}^{q'-1} \mathbb{P}(E_{k,2}(\tau_{k,q},j) \mid \mathcal{H}_{\tau_{k,q},j}) \right],
\]

here \( \tau_{k,q,j} \) is the round when the event \( E_{k,1}(t) \) has happened for the \( j \)-th time after round \( \tau_{k,q} + 1 \).

The right hand side of the above equality is the expectation of a time-varying geometric distribution with the success probability of the \( j \)-th trial being \( \mathbb{P}(E_{k,2}(\tau_{k,q},j) \mid \mathcal{H}_{\tau_{k,q},j}) \). We can lower bound this probability by

\[
\inf_{t \geq \tau_{k,q}+1} \mathbb{P}(E_{k,2}(t) \mid \mathcal{H}_t) = \inf_{t \geq \tau_{k,q}+1} \prod_{i \in s_{g,k}} \mathbb{P}(\|\theta_{t,i} - \mu_i\| \leq \varepsilon \mid \mathcal{H}_t).
\]

Then according to the monotonicity of the expectation, we can upper bound the above expectation by

\[
\mathbb{E} \left[ \sum_{q \geq 1} (q' - 1) \mathbb{P}(E_{k,2}(\tau_{k,q},q') \mid \mathcal{H}_{\tau_{k,q},q'}) \prod_{j=1}^{q'-1} \mathbb{P}(E_{k,2}(\tau_{k,q},j) \mid \mathcal{H}_{\tau_{k,q},j}) \right]
\]
\[
\text{Lemma 3. In Algorithm 2, for any } k \in [K], \text{ we have}
\]
\[
E \left[ \sum_{t=1}^{T} 1 \{ s_{t,k} = s_{g,k} \mid \| \theta_{t,s_{g,k}} - \mu_{s_{g,k}} \|_\infty > \varepsilon \} \right] \leq |s_{g,k}| \left( 2 + \frac{8}{\varepsilon^2} \right).
\]

\text{Proof. We first decompose the event as}
\[
E \left[ \sum_{t=1}^{T} 1 \{ s_{t,k} = s_{g,k} \mid \| \theta_{t,s_{g,k}} - \mu_{s_{g,k}} \|_\infty > \varepsilon \} \right] \\
\leq E \left[ \sum_{t=1}^{T} 1 \{ s_{t,k} = s_{g,k} \mid \| \hat{\mu}_{t,s_{g,k}} - \mu_{s_{g,k}} \|_\infty > \varepsilon / 2 \} \right] \\
+ E \left[ \sum_{t=1}^{T} 1 \{ s_{t,k} = s_{g,k} \mid \| \theta_{t,s_{g,k}} - \hat{\mu}_{t,s_{g,k}} \|_\infty > \varepsilon / 2 \} \right].
\]

For term (16), we have
\[
(16) = E \left[ \sum_{t=1}^{T} 1 \{ s_{t,k} = s_{g,k}, \exists i \in s_{g,k} : |\hat{\mu}_{t,i} - \mu_i| > \varepsilon / 2 \} \right] \\
\leq \sum_{i \in s_{g,k}} E \left[ \sum_{t=1}^{T} 1 \{ i \in \cup S_t : |\hat{\mu}_{t,i} - \mu_i| > \varepsilon / 2 \} \right] \\
= \sum_{i \in s_{g,k}} \sum_{w=0}^{T} E \left[ \sum_{t \in [T]: N_{t,i} = w} 1 \{ i \in \cup S_t : |\hat{\mu}_{t,i} - \mu_i| > \varepsilon / 2 \} \right] \\
\leq \sum_{i \in s_{g,k}} \sum_{w=0}^{T} P \left( |\hat{\mu}_{t,i} - \mu_i| > \varepsilon / 2 \text{ for } t \text{ satisfies } N_{t,i} = w \text{ and } N_{t+1,i} = w + 1 \right) \\
\leq \sum_{i \in s_{g,k}} \left( 1 + \sum_{w=1}^{T} P \left( |\hat{\mu}_{t,i} - \mu_i| > \varepsilon / 2 \text{ for } t \text{ satisfies } N_{t,i} = w \text{ and } N_{t+1,i} = w + 1 \right) \right) \\
\leq \sum_{i \in s_{g,k}} \left( 1 + 2 \sum_{w=1}^{T} \exp \left( -\frac{w \varepsilon^2}{2} \right) \right) \\
\leq \left| s_{g,k} \right| \left( 1 + 2 \sum_{w=1}^{\infty} \left( \exp \left( -\frac{\varepsilon^2}{2} \right) \right)^w \right) \\
\leq \left| s_{g,k} \right| \left( 1 + 2 \exp \left( -\frac{\varepsilon^2}{2} \right) \right) \\
\leq \left| s_{g,k} \right| \left( 1 + \frac{4}{\varepsilon^2} \right),
\]

where (18) is due to the Lemma 5. Similarly, we have the following bound for term (17),
\[
(17) = E \left[ \sum_{t=1}^{T} 1 \{ s_{t,k} = s_{g,k}, \exists i \in s_{g,k} : |\hat{\mu}_{t,i} - \mu_i| > \varepsilon / 2 \} \right]
\]
\[
\leq \sum_{i \in s_{g,k}} E \left[ \sum_{t=1}^{T} \mathbf{1} \{ i \in \cup S_t : |\hat{\mu}_{t,i} - \theta_{t,i}| > \frac{\varepsilon}{2} \} \right]
\]
\[
= \sum_{i \in s_{g,k}} E \left[ \sum_{w=0}^{T} E \left[ \sum_{t \in [T] : N_{t,i} = w} \mathbf{1} \{ i \in \cup S_t : |\hat{\mu}_{t,i} - \theta_{t,i}| > \frac{\varepsilon}{2} \} \right] \right]
\]
\[
\leq \sum_{i \in s_{g,k}} E \left[ \sum_{w=0}^{T} \mathbb{P} \left( |\hat{\mu}_{t,i} - \theta_{t,i}| > \frac{\varepsilon}{2} \text{ for } t \text{ satisfies } N_{t,i} = w \text{ and } N_{t+1,i} = w + 1 \right) \right]
\]
\[
\leq \sum_{i \in s_{g,k}} \left( 1 + \sum_{w=1}^{T} \exp \left( -\frac{w \varepsilon^2}{2} \right) \right)
\]
\[
\leq |s_{g,k}| \left( 1 + \sum_{w=1}^{\infty} \left( \exp \left( -\frac{\varepsilon^2}{2} \right) \right)^w \right)
\]
\[
\leq |s_{g,k}| \left( 1 + 2 \exp \left( -\frac{\varepsilon^2}{2} \right) \right)
\]
\[
\leq |s_{g,k}| \left( 1 + 4 \frac{\varepsilon^2}{2} \right),
\]
where (19) is due to the Lemma 7.

Above all, we have

\[
(16) + (17) \leq |s_{g,k}| \left( 1 + 4 \frac{\varepsilon^2}{2} \right) + |s_{g,k}| \left( 1 + 4 \frac{\varepsilon^2}{2} \right)
\]
\[
= |s_{g,k}| \left( 2 + 8 \frac{\varepsilon^2}{2} \right).
\]

\[\square\]

D Analysis of CTS with Gaussian Priors and Greedy Oracle

**Theorem 3.** When the reward distribution for each arm \( i \) is \( D_i = \mathcal{N}(\mu_i, 1) \), the cumulative greedy regret of Algorithm 3 can be upper bounded by

\[
R_g(T) \leq \sum_{s \neq s_{g,1}} \max_{k : s \notin S_{g,k}} \frac{B \varepsilon^2 |s|^2 \Delta_{s,k} \log T}{2B |\cup S_g|} + \sum_{k \in [K]} \frac{C'}{\varepsilon^2} \left| s_{g,k} \right| \Delta_{s,k} \max
\]
\[
+ \left( |\cup S_g| \left( 2 + \frac{8 \varepsilon^2}{2} \right) + m \right) \Delta_{s,k} \max
\]
\[
= O \left( \sum_{s \neq s_{g,1}} \max_{k : s \notin S_{g,k}} \frac{B \varepsilon^2 |s|^2 \Delta_{s,k} \log T}{\Delta_{s,k}^2} \right),
\]

for any \( \varepsilon \) such that \( \forall s \neq s_{g,1} \) and \( k \) satisfying \( s \notin S_{g,k}, \Delta_{s,k} > 2B |\cup S_g| \varepsilon \), where \( B \) is the coefficient of the Lipschitz continuity condition, \( |\cup S_g| \) is the number of base arms that belong to the units contained in \( S_g \), \( C \) and \( C' \) are two universal constants.

**Proof.** The proof of Theorem 3 is very similar to that of Theorem 2 in Appendix C, here we only state the differences. When proving Theorem 3, we redefine \( L(s), B(t) \) and \( C(t) \) with slight differences.
in constants as

\[ L(s) = \max_{k : s \notin \mathcal{S}_{s,k}} \frac{8B^2 |s|^2 \log T}{(\Delta_{s,k} - 2B |\cup \mathcal{S}_g| \varepsilon)^2}, \]

\[ B(t) = \left\{ \exists i \in [m] : |\theta_{t,i} - \mu_{t,i}| > \sqrt{\frac{2 \log T}{N_{t,i}}} \right\}, \]

\[ C(t) = \left\{ \exists i \in [m] : |\mu_{t,i} - \mu_i| > \sqrt{\frac{2 \log T}{N_{t,i}}} \right\}. \]

Based on these new definitions, we use the result of Lemma 6 instead of the result of Lemma 5 and 7 to get an upper bound for Eq (10) and (11). Besides, when proving Lemma 1, we use the upper bound for [24, Eq(4)] instead of using the result of Lemma 8.

\[ \square \]

E  Regret Analysis for the Case of Multiple Greedy Solutions under \( \mu \)

In this section, we discuss how to extend the proof of Theorem 2 to the case where multiple solutions can be returned by Greedy with input \( \mu \), or equivalently the optimal unit in each step \( k \) (Line 4 in Algorithm 1) may not be unique.

For any \( k \in [K] \), define

\[ \sigma_k = \left\{ s_1, s_2, \ldots, s_k : s_1 \in \arg\max_s r\left( \{s\} \cup \{s_i \mid i \neq 1\} \right), \ldots, s_k \in \arg\max_s r\left( \{s_1, \ldots, s_{k-1}, s\} \right) \right\} \]

as the set of actions containing \( k \) units that could be selected by Greedy in the first \( k \) steps. And denote

\[ \bar{\sigma}_k = \{ s_k : s_1, s_2, \ldots, s_k \in \sigma_k \} \]

as the set of units which may be selected as the \( k \)-th unit by Greedy under \( \mu \). Let \( \sigma_0 = \emptyset \). Then for any \( k \) and \( S \in \sigma_{k-1} \), define

\[ \sigma_k(S) = \left\{ s : s \in \arg\max_{s \notin S} r(S \cup \{s\}) \right\} \]

as the set of units that may be selected by Greedy in the \( k \)-th step when \( S \) is selected in the previous \( k-1 \) steps. Let \( \sigma_1(\emptyset) = \{ s : s \in \arg\max_{s \in \mathcal{S}_1} r(\{s\}) \} \). Denote \( s_{g,s_k} \in \arg\min_{s \in \sigma_k(S)} |s| \) as one of the optimal \( k \)-th unit when \( S \) is selected with the minimum unit size. And for any \( s \notin S \cup \sigma_k(S) \), define \( \Delta_{s,S,k} = r(S \cup \{s_{g,s_k}\}) - r(S \cup \{s\}) \) as the corresponding reward gap. Let \( S'_g \in \arg\min_{S \in \sigma_k} |S| \) be one of the possible actions returned by Greedy containing the maximum number of base arms. Then for any unit \( s \), define

\[ L(s) = \max_{k \in[K], S \in \sigma_{k-1} : s \notin \sigma_k(S)} \frac{6B^2 |s|^2 \log T}{(\Delta_{s,S,k} - 2B |\cup \mathcal{S}_g| \varepsilon)^2}. \]

Note \( L(s) = 0 \) if \( \{ k \in [K], S \in \sigma_{k-1} : s \notin \sigma_k(S) \} = \emptyset \).

In this case, we regard how Greedy breaks the tie at each step as a black box. In order to take into account the worst case where Greedy always return a solution with minimum reward compared to other possible solutions under \( \mu \), we define \( S_g \in \arg\min_{S \in \sigma_k} r(S, \mu) \) as one of the possible actions returned by Greedy with the minimum expected reward, and define the cumulative greedy regret as

\[ R_g(T) = \mathbb{E} \left[ \sum_{t=1}^T \Delta_{S_t} \right] = \mathbb{E} \left[ \sum_{t=1}^T \max \{ r(S_g, \mu) - r(S_t, \mu), 0 \} \right]. \]

**Theorem 4.** When there are multiple Greedy solutions with input \( \mu \), the cumulative greedy regret of Algorithm 2 can be bounded by

\[ R_g(T) \leq \sum_{s} \max_{k \in[K], S \in \sigma_{k-1} : s \notin \sigma_k(S)} \frac{6B^2 |s|^2 \Delta_{s,s_k} \log T}{(\Delta_{s,S,k} - 2B |\cup \mathcal{S}_g| \varepsilon)^2}. \]

23
for any $\varepsilon$ such that $\forall k \in [K], S \in \sigma_{k-1}, s \notin \sigma_k(S)$, there is $\Delta_{s,S,k} > 2B \| \cup S' \| \varepsilon$, where $B$ is the coefficient of the Lipschitz continuity and $| \cup S |$ is the number of base arms that belong to the units contained in $S$, $C$ and $C'$ are two universal constants.

Proof. With the same definition of $B(t)$ and $C(t)$ as in Section C, the greedy regret can be decomposed by

$$R_g(T) \leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{B(t)\} \Delta_{S_t} \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{C(t)\} \Delta_{S_t} \right] \leq 4m \Delta_{\max}.$$ (20)

Same as the proof for the second and the third term in (9), we have the following bound for the first and the second term in (20),

$$\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{B(t)\} \Delta_{S_t} \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{C(t)\} \Delta_{S_t} \right] \leq 4m \Delta_{\max}.$$ (20)

The main difference is to bound the last term. Intuitively, if $S_t \in \sigma_K$, we would have $\Delta_{S_t} = 0$. Thus we only need to consider the case where $S_t \notin \sigma_K$ to bound the regret. We will analyze such case by sequentially analyzing the $s_{t,k}$ for each $k = 1, 2, \ldots, K$. According to this idea, the first term can be bounded by

$$\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{-B(t), -C(t)\} \Delta_{S_t} \right] \leq \sum_{k \in [K]} \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{-B(t), -C(t), S_{t,k-1} \in \sigma_{k-1}, \| \theta_{t,S_{t,k-1}} - \mu_{S_{t,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}) \} \Delta_{S_t} \right] \\
+ \sum_{k \in [K]} \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{s_{t,k} \in \sigma_k, \| \theta_{t,s_k} - \mu_{s_k} \|_\infty > \varepsilon \} \right] \Delta_{\max}.$$

For the first term, According to Lemma 4, we have

$$\sum_{k \in [K]} \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{-B(t), -C(t), S_{t,k-1} \in \sigma_{k-1}, \| \theta_{t,S_{t,k-1}} - \mu_{S_{t,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}) \} \Delta_{S_t} \right] \leq \sum_{k \in [K]} \sum_{s} \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1}\{s = s_{t,k}, N_{t,s} \leq L(s) \} \Delta_{S_t} \right] + \sum_{k \in [K]} \sum_{S \in \sigma_{k-1}} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{t,k}, s|} \Delta_{\max}$$

$$\leq \mathbb{E} \left[ \sum_{s} \sum_{t} \mathbb{1}\{s \in S_t, N_{t,s} \leq L(s) \} \Delta_{S_t} \right] + \sum_{k \in [K]} \sum_{S \in \sigma_{k-1}} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{t,k}, s|} \Delta_{\max}$$

$$\leq \sum_{s} \max_{s \in [K]} \sum_{s \in \sigma_{k-1}} \frac{6B^2 \| s \|^2 \Delta_{\max} \log T}{\Delta_{s,S,k} - 2B | \cup S' | \varepsilon^2} + \sum_{k \in [K]} \sum_{S \in \sigma_{k-1}} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{|s_{t,k}, s|} \Delta_{\max},$$

where $C, C'$ are two universal constants.
For the second term, we have
\[
\sum_{k \in [K]} \mathbb{E} \left[ \sum_{l=1}^{T} \mathbb{1} \{ s_{l,k} \in s_k, \| \theta_{t,s_{l,k}} - \mu_{s_{l,k}} \|_\infty > \varepsilon \} \right] \leq \sum_{k \in [K]} \sum_{s \in s_k} \mathbb{E} \left[ \mathbb{1} \{ s_{l,k} = s, \| \theta_{t,s} - \mu_{s} \|_\infty > \varepsilon \} \right] \\
\leq \sum_{k \in [K]} \sum_{s \in s_k} |s| \left( 2 + \frac{8}{\varepsilon^2} \right),
\]
where the last inequality is obtained by applying the result of Lemma 3.

Above all, we have the following upper bound for the greedy regret
\[
R_g(T) \leq \sum_{s} \max_{k \in [K], s \in s_{k-1}, s \notin \sigma_k(S)} B \mathbb{E} \left[ |s| \right] 25 \log T \left( \Delta_{s,S,k} - 2B \left| \cup S' \right| \varepsilon \right)^2 \\
+ \left( 4m + \sum_{k \in [K]} \sum_{s \in s_k} |s| \left( 2 + \frac{8}{\varepsilon^2} \right) + \sum_{k \in [K]} \sum_{s \in s_k} \frac{C}{\varepsilon} \left( \frac{C'}{\varepsilon} \right) \right) \Delta_{\text{max}},
\]

Lemma 4. For Algorithm 2, when there are multiple Greedy solutions with input \( \mu \), for any \( k \in [K] \), we have
\[
\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1} \{ -B(t), -C(t), S_{t,k-1} \in \sigma_{k-1}, \| \theta_{t,S_{t,k-1}} - \mu_{S_{t,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}) \} \Delta_{S_t} \right] \\
\leq \mathbb{E} \left[ \sum_{s} \sum_{t=1}^{T} \mathbb{1} \{ s = s_{t,k}, N_{t,s} \leq L(s) \} \Delta_{S_t} \right] + \sum_{S \in \sigma_{k-1}} \frac{C}{\varepsilon} \left( \frac{C'}{\varepsilon} \right) \Delta_{\text{max}},
\]
where \( C, C' \) are two universal constants.

Proof. Recall given \( S \in \sigma_{k-1} \), for any unit \( s \notin S \cup \sigma_k(S) \), \( \Delta_{S,S,k} = r(S \cup \{ s_{g,s,k} \}, \mu) - r(S \cup \{ s \}, \mu) \). Define the event
\[
D_k(t) = \left\{ B \sum_{i \in S_{t,k}} |\theta_{t,i} - \mu_i| > \Delta_{S_{t,k},S_{t,k-1},k} - B \left( 2 \left| \cup S_{t,k-1} \right| + \left| s_{g,S_{t,k-1},k} \right| + 1 \right) \varepsilon \right\}.
\]
Then the formula in Lemma 1 can further bounded by
\[
\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1} \{ -B(t), -C(t), S_{t,k-1} \in \sigma_{k-1}, \| \theta_{t,S_{t,k-1}} - \mu_{S_{t,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}) \} \right] \\
\leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1} \{ -B(t), -C(t), S_{t,k-1} \in \sigma_{k-1}, \| \theta_{t,S_{t,k-1}} - \mu_{S_{t,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}), D_k(t) \} \Delta_{S_t} \right] \\
+ \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1} \{ -B(t), -C(t), S_{t,k-1} \in \sigma_{k-1}, \| \theta_{t,S_{t,k-1}} - \mu_{S_{t,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}), -D_k(t) \} \Delta_{\text{max}} \right].
\]
(21)
(22)

For term (21), we claim that the event
\[
\{ -B(t), -C(t), S_{t,k-1} \in \sigma_{k-1}, \| \theta_{t,S_{t,k-1}} - \mu_{S_{t,k-1}} \|_\infty \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}), D_k(t) \}
\]
implies \( N_{t,s_{t,k}} \leq L(s_{t,k}) \). This claim can be proved by contradiction.

Suppose \( N_{t,s_{t,k}} > L(s_{t,k}) \), then we must have
\[
B \sum_{i \in s_{t,k}} |\theta_{t,i} - \mu_i| \leq B \sum_{i \in s_{t,k}} \sqrt{\frac{6 \log T}{N_{t,s_{t,k}}}}.
\]
We now prove that for any

We claim that if the event

ε


For term (22), we first define event

E

happens, then

Next we prove that for any unit

in the first

r

′

(S,k, ∪

\{−1 \leq k < s \leq 1\}) = \log T \left(\frac{6 \log T}{6B^2 |s_t,k|^2 \log T} \left(\Delta_{s_t,k,S_t,k-1,k} - B \left|\cup S'_g \right| \varepsilon\right) \right.

\leq B |s_t,k| \left|\frac{\log T}{B^2 |s_t,k|^2 \log T} \left(\Delta_{s_t,k,S_t,k-1,k} - B \left(2 \left|\cup S_{t,k-1} \right| + |s_g,S_t,k-1,k| + 1\right) \varepsilon\right) \right.

= \Delta_{s_t,k,S_t,k-1,k} - B \left(2 \left|\cup S_{t,k-1} \right| + |s_g,S_t,k-1,k| + 1\right) \varepsilon,

where the first inequality is due to the event of \(-B(t)\) and \(-C(t)\), the second inequality is obtained by substituting \(N_{t,s,k} \) with \(L(s_t,k)\) and the third one comes from the definition of \(S'_g\) and the fact that \(S_{t,k-1} \in \sigma_{k-1}\). Thus we conclude the event \(D_{t,k}(t)\) will not happen and the claim is proved.

Then according to the above claim, there is

\[
(21) \leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I} \left\{ (N_{t,s,k} \leq L(s_t,k), S_{t,k-1} \in \sigma_{k-1}, s_t,k \notin \sigma_k(S_{t,k-1}) \} \right\} \Delta_{s_t,k} \right]

\leq \mathbb{E} \left[ \sum_{s} \sum_{t=1}^{T} \mathbb{I} \left\{ s = s_t,k, N_{t,s} \leq L(s) \right\} \Delta_{s_t,k} \right].

For term (22), we first define event \(E_{S,k,1}(t)\) for \(S \in \sigma_{k-1}\) as

\[E_{S,k,1}(t) = \{ \forall \theta' \text{ with } \theta'_i = \theta_{i,j} \text{ for any } i \notin s_g,S_{k,k} \text{ and } \|\theta'_{s_g,S_{k,k}} - \mu_{s_g,S_{k,k}}\|_\infty \leq \varepsilon, \text{ then } s_g,S_{k,k} \text{ is the } k\text{-th selected unit by Greedy when the input is } \theta' \} \]

and the event \(E_{S,k,2}(t)\) as

\[E_{S,k,2}(t) = \{ \|\theta_{s_g,S_{k,k}} - \mu_{s_g,S_{k,k}}\|_\infty > \varepsilon \}. \]

We claim that if the event \(\{ S_{t,k-1} \in \sigma_{k-1}, \|\theta_{s_g,S_{k,k}} - \mu_{s_g,S_{k,k}}\|_\infty \leq \varepsilon, s_t,k \notin \sigma_k(S_{t,k-1}), -D_{t,k}(t) \}\) happens, then \(E_{S_{t,k-1},k,1}(t)\) and \(E_{S_{t,k-1},k,2}(t)\) hold.

To prove \(E_{S_{t,k-1},k,1}(t)\) holds, it is sufficient to prove for any \(\theta'\) defined in \(E_{S_{t,k-1},k,1}(t)\), \(S_{t,k-1}\) is still the set of units selected by Greedy in the first \(k-1\) steps with input \(\theta'\) and for any unit \(s' \notin S_{t,k} \cup \{ s_g,S_{t,k-1} \}\), \(r(S_{t,k-1} \cup \{ s' \}, \theta') < r(S_{t,k-1} \cup \{ s_g,S_{t,k-1} \}, \theta')\) holds.

We now prove that for any \(\theta'\) satisfying the condition defined in \(E_{S_{t,k-1},k,1}(t)\), \(S_{t,k-1}\) is still the set of units selected by Greedy in the first \(k-1\) steps with input \(\theta'\). The event \(S_{t,k-1}\) is selected in the first \(k-1\) steps with input \(\theta_1\) means that for any \(k' < k, s \notin S_{t,k'}, r(S_{t,k'}, \theta_1) > r(S_{t,k'-1} \cup \{ s \}, \theta_1)\). The mean vector \(\theta'\) and \(\theta_1\) are only different on \(s_g,S_{t,k-1}\), thus for any \(k' < k, s \notin S_{t,k'} \cup \{ s_g,S_{t,k-1} \}\), we still have \(r(S_{t,k'}, \theta') > r(S_{t,k'-1} \cup \{ s \}, \theta')\). And for the unit \(s_g,S_{t,k-1},k\), for any \(k' < k\), we have

\[
r(S_{t,k'}, \theta') = r(S_{t,k'}, \theta_1) \geq r(S_{t,k'-1} \cup \{ s_g,S_{t,k-1} \}, \mu) + \Delta_{s_g,S_{t,k-1},k,S_{t,k'-1},k'} - B |\cup S_{t,k'}| \varepsilon
\]

\[
\geq r(S_{t,k'-1} \cup \{ s_g,S_{t,k-1} \}, \theta_1) - B \left( |\cup S_{t,k'}| + |\cup S_{t,k'-1}| + |s_g,S_{t,k-1,k}| \right) \varepsilon + \Delta_{s_g,S_{t,k-1},k,S_{t,k'-1},k'}
\]

\[
\geq r(S_{t,k'-1} \cup \{ s_g,S_{t,k-1} \}, \theta') + \Delta_{s_g,S_{t,k-1},k,S_{t,k'-1},k'} - 2B |S'_g| \varepsilon
\]

The last inequality holds due to the requirement of \(\varepsilon\) in Theorem 4. Above all, we conclude \(\forall k' < k, s \notin S_{t,k'}, r(S_{t,k'}, \theta') > r(S_{t,k'-1} \cup \{ s \}, \theta')\), thus \(S_{t,k-1}\) is still the set of units selected by Greedy in the first \(k-1\) steps under \(\theta'\).

Next we prove that for any unit \(s' \notin S_{t,k-1} \cup \{ s_g,S_{t,k-1} \}\), \(r(S_{t,k-1} \cup \{ s' \}, \theta') < r(S_{t,k-1} \cup \{ s_g,S_{t,k-1} \}, \theta')\) holds.

\[
r(S_{t,k-1} \cup \{ s' \}, \theta') = r(S_{t,k-1} \cup \{ s' \}, \theta_1)
\]
\[
\leq r(S_{t,k}, \theta_t), \text{ (Greedy's property)}
\]
\[
\leq r(S_{t,k}, \mu) + B \sum_{i \in \mathcal{S}_{t,k-1}} |\theta_{t,i} - \mu_i| + B \sum_{i \in \mathcal{S}_{t,k}} |\theta_{t,i} - \mu_i| \quad \text{(Lipschitz continuity)}
\]
\[
\leq r(S_{t,k}, \mu) + B |\mathcal{S}_{t,k-1}| \varepsilon + B \sum_{i \in \mathcal{S}_{t,k}} |\theta_{t,i} - \mu_i|
\]
\[
\leq r(S_{t,k}, \mu) + B |\mathcal{S}_{t,k-1}| \varepsilon + \Delta_{s_{t,k},s_{t,k-1},k} - B (2|\mathcal{S}_{t,k-1}| + |s_{g,s_{t,k-1},k}| + 1) \varepsilon
\]
\[
\leq r(S_{t,k-1} \cup \{s_{g,s_{t,k-1},k}\}, \mu) + B |\mathcal{S}_{t,k-1}| \varepsilon - B (2|\mathcal{S}_{t,k-1}| + |s_{g,s_{t,k-1},k}| + 1) \varepsilon
\]
\[
= r(S_{t,k-1} \cup \{s_{g,s_{t,k-1},k}\}, \mu) - B (|\mathcal{S}_{t,k-1}| + |s_{g,s_{t,k-1},k}| + 1) \varepsilon
\]
\[
\leq r(S_{t,k-1} \cup \{s_{g,s_{t,k-1},k}\}, \theta') - B \varepsilon \quad \text{(Lipschitz continuity)}
\]
\[
< r(S_{t,k-1} \cup \{s_{g,s_{t,k-1},k}\}, \theta'),
\]
where the first equality is because \(\theta'_i = \delta_{t,i}\) for any \(i \in s'\) and \(i \in \mathcal{S}_{t,k-1}\), the third inequality is because \(\|\theta_{t,s_{t,k-1}} - \mu_{s_{t,k-1}}\|_{\infty} \leq \varepsilon\). (23) comes from the definition of \(-D_k(t)\) and (24) is due to the definition of \(\Delta_{s_{t,k},s_{t,k-1},k}\).

Above all, we have proved if event \(\{S_{t,k-1} \in \sigma_{k-1}, \|\theta_{t,s_{t,k-1}} - \mu_{s_{t,k-1}}\|_{\infty} \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}), -D_k(t)\}\) happens, then \(E_{s_{t,k-1},k,1}(t)\) holds.

Next we consider \(E_{s_{t,k-1},k,2}(t)\). By contradiction, when the event \(\{S_{t,k-1} \in \sigma_{k-1}, \|\theta_{t,s_{t,k-1}} - \mu_{s_{t,k-1}}\|_{\infty} \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}), -D_k(t)\}\) happens, if \(-E_{s_{t,k-1},k,1}(t) = \{\|\theta_{t,s_{t,k-1}} - \mu_{s_{t,k-1}}\|_{\infty} \leq \varepsilon\}\) holds, then \(\theta_t\) satisfies the property of \(\theta'\) defined in \(E_{s_{t,k-1},k,1}(t)\). Thus according to \(E_{s_{t,k-1},k,1}(t), s_{g,s_{t,k-1},k}\) would be the \(k\)-th selected unit by Greedy when the input is \(\theta_t\), or in other words \(s_{t,k} = s_{g,s_{t,k-1},k} \in \sigma_k(S_{t,k-1})\). This contradicts \(\{S_{t,k-1} \in \sigma_{k-1}, \|\theta_{t,s_{t,k-1}} - \mu_{s_{t,k-1}}\|_{\infty} \leq \varepsilon, s_{t,k} \notin \sigma_k(S_{t,k-1}), -D_k(t)\}\). Thus we conclude \(E_{s_{t,k-1},k,2}(t)\) also holds.

Above all, for term (22) we have
\[
(22) \leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbbm{1}\{S_{t,k-1} \in \sigma_{k-1}, E_{s_{t,k-1},k,1}(t), E_{s_{t,k-1},k,2}(t)\} \right]
\]
\[
\leq \sum_{S \in \sigma_{k-1}} \mathbb{E} \left[ \sum_{t=1}^{T} \mathbbm{1}\{E_{S,k,1}(t), E_{S,k,2}(t)\} \right]
\]
\[
\leq \sum_{S \in \sigma_{k-1}} \sum_{q \geq 0} \mathbb{E} \left[ \sum_{t=\tau_{S,q+1}}^{\tau_{S,q}} \mathbbm{1}\{E_{S,k,1}(t), E_{S,k,2}(t)\} \right]
\]
\[
\leq \sum_{S \in \sigma_{k-1}} \sum_{q \geq 0} \mathbb{E} \left[ \sup_{t \geq \tau_{S,q+1}} \prod_{i \in s_{g,r,k}} \frac{1}{P(|\theta_{t,i} - \mu_i| \leq \varepsilon \mid x_i)} - 1 \right]
\]
\[
\leq \sum_{S \in \sigma_{k-1}} \left( \sum_{q=0}^{[8/\varepsilon^2]} (\varepsilon^{-4})^{[s_{g,s,k}]} + \sum_{q \geq [8/\varepsilon^2]} e^{-\varepsilon^2 q/8} (\varepsilon^{-4})^{[s_{g,r,k}]} \right)
\]
\[
\leq \sum_{S \in \sigma_{k-1}} \frac{C}{\varepsilon^2} \left( \frac{C'}{\varepsilon^4} \right)^{[s_{g,s,k}]} ,
\]
where \(\tau_{S,q}\) is the round at which \(E_{S,k,1}(t) \land -E_{S,k,2}(t)\) occurs for the \(q\)-th time, note \(\tau_{S,0} = 0\) for any \(k \in [K], S \in \sigma_{k-1}\). The fourth inequality comes from the result in Lemma 2 and the fifth comes from the Lemma 8. Here \(C, C'\) are two universal constants.
F Technical Lemmas

Lemma 5. (Chernoff-Hoeffding bound) Let \(X_1, X_2, \ldots, X_n\) be identical independent random variables such that \(X_i \in [0, 1]\) and \(\mathbb{E}[X_i] = \mu\) for any \(i \in [n]\). Then for any \(\epsilon \geq 0\), we have

\[
\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2 \exp\left(-2n\epsilon^2\right)
\]

Lemma 6. (Concentration and anti-concentration inequalities for Gaussian distributed random variables [11].) For a Gaussian distributed random variable \(Z\) with mean \(m\) and variance \(\sigma^2\), for any \(z\),

\[
\frac{1}{4\sqrt{n}} \exp\left(-\frac{7z^2}{2}\right) < \mathbb{P}(|Z - m| > z\sigma) \leq \frac{1}{2} \exp\left(-\frac{z^2}{2}\right).
\]

Lemma 7. (Lemma 3 in [30]) In Algorithm 2, for any base arm \(i \in [m]\) and round \(t\), we have

\[
\mathbb{P}(|\theta_{t,i} - \hat{\mu}_{t,i}| > \epsilon | a_{t,i}, b_{t,i}) \leq 2 \exp\left(-2N_{t,i}\epsilon^2\right),
\]

where \(a_{t,i}, b_{t,i}\) are the value of \(a_i\) and \(b_i\) before the start of round \(t\).

Lemma 8. (Lemma 5 in [24]) In Algorithm 2, for any unit \(s\), let \(\tau'_q = \min\{t : N_{t,s} \geq q\}\), we have

\[
\mathbb{E}\left[\sup_{t \geq \tau'_q} \prod_{i \in s} \mathbb{P}(|\theta_{t,i} - \mu_i| \leq \epsilon | H_t)\right] - 1 \leq \begin{cases} (c\epsilon^{-4})^{|s|} & \text{for every } q \geq 0 \\ e^{-\epsilon q/8} (c'\epsilon^{-4})^{|s|} & \text{if } q \geq 8/\epsilon^2, \end{cases}
\]

where \(c\) and \(c'\) are two universal constants.