Exploiting Correlation in Finite-Armed Structured Bandits

Samarth Gupta
Carnegie Mellon University

Shreyas Chaudhari
IIT Madras

Gauri Joshi
Carnegie Mellon University

Osman Yağan
Carnegie Mellon University

Abstract

We consider a structured Multi-Armed bandit problem in which mean rewards of different arms are related through a hidden parameter. We propose an approach that allows generalization of classical bandit algorithms such as UCB and Thompson sampling to the structured bandit setting. Our approach is based on exploiting the structure in the problem to identify some arms as sub-optimal and pulling them only $O(1)$ times. This results in significant reduction in cumulative regret, and in fact our algorithm achieves bounded (i.e., $O(1)$) regret whenever possible. We empirically demonstrate the superiority of our algorithms via simulations and experiments on the Movielens dataset. Moreover, the problem setting we study in this paper subsumes several previously studied framework such as Global, Regional and Structured bandits with linear rewards.

1 Introduction

The Multi-armed bandit problem (MAB) falls under the umbrella of sequential decision-making problems and is widely studied because of its numerous applications in medical diagnosis, system testing, scheduling in computing systems, and web optimization etc. In the classical $K$-armed bandit formulation, a player is presented with $K$ arms. At each time step $t$, she decides to pull an arm $k \in K$ and receives a random reward $R_t$ with unknown mean $\mu_k$. The goal of the player is to maximize their cumulative reward. The seminal work of Lai and Robbins [1] proposed the UCB (upper confidence bound algorithm) that balances the exploration-exploitation trade-off in the MAB problem. Subsequently, several other algorithms such as Thompson Sampling [6] and KL-UCB [7] were proposed and analyzed for the classical MAB setting.

In this work, we study a setting in which rewards corresponding to different arms are related to each other through a hidden parameter $\theta$; see Section 2.1 and Figure 1. Our main goal is to leverage these relations to reduce the amount of exploration and thus achieve a cumulative reward that is significantly higher than possible in the classical MAB setting.

There are many practical applications where the model studied here can be useful. For instance, let us consider the example of ad selection, where a company needs to decide which version of the ad it needs to display to the user. It has different versions for the same ad and depending on which ad is displayed the user engagement (in terms of click probability and time spent looking at the ad) is affected. In order to maximize user engagement, the company needs to identify the most appealing ad for the user in an online manner and this is where multi-armed bandit algorithms like UCB, Thompson Sampling and KL-UCB can be helpful. However, classical MAB algorithms are typically based on the assumption that rewards corresponding to all arms are independent of each other. This assumption is unlikely to hold in reality since the user choices corresponding to different versions of an ad are likely to be related to each other; e.g., the choices corresponding to different versions may depend on the age/occupation/income of the user.

Alternatively, contextual bandits [8] can be considered, where the player also observes the context feature of the user to whom ad is displayed (i.e., their age/occupation/income information). By trying to learn a mapping from feature information to the most appealing arm, contextual bandit algorithms prove useful for the application of targeted advertising. However, observing contextual features leads to privacy concerns. Also, the contextual features may not be visible for the users that are signed in anonymously. This raises the question of
whether it is possible to do targeted advertising for a user without observing their features.

Apart from ad selection, there are many other applications where the structured bandit problem described above can be useful. For instance, in the dynamic pricing problem [2], a player needs to select the price of a product from a finite set of prices \( P \), and the average revenue in time slot \( t \) is a function of the selected price \( p_t \) and the market size \( \theta \). These functions are typically known from literature [10], but the pricing decisions \( p_t \) need to be made without knowing the market size such that the total revenue is maximized; hence, this problem fits perfectly in our setting. Other applications of this model include cellular coverage optimization [11], drug dosage optimization [12] and System diagnosis (See [11][12] for a description on how \( \mu_k(\theta) \) are known through literature in these cases). Our general treatment of the structured bandit setting allows our work to be helpful in all these problems.

Motivated by these, we consider a structured multi-armed bandit problem in which the mean reward of different arms are related through a common hidden parameter \( \theta \). Specifically, the expected reward of arm \( k \) is given as \( \mathbb{E}[R_k|\theta] = \mu_k(\theta) \). In the setting considered, the mean reward functions, \( \mu_k(\theta) \), are known to the player but the true value of shared parameter \( \theta^* \) is unknown. The dependence on the common parameter introduces a structure in this MAB problem – the rewards observed from an arm can provide information about mean rewards from other arms. In the aforementioned example of ad selection, the mean reward mappings from user contexts to different versions of ad can be learned from paid surveys in which users participate with their consent. Such structured bandit models have been studied in the past [2][12][13][14][15][16], but with certain restriction on the mean reward mappings \( \mu_k(\theta) \). In this work, we allow a general treatment of the problem that poses no restriction on the mean reward mappings and in fact subsumes the models studied before; see [2][2] for details.

**Main Contributions.**

1) We consider a general setting for the structured multi-armed problem which subsumes recently studied models such as the Global Bandits [9], Regional Bandits [12] and Structured bandits with linear functions [13].

2) We develop a novel approach that exploits the structure of the bandit problem to identify *sub-optimal* arms. In particular, we generate an estimate of \( \theta \) at each round to identify competitive and non-competitive arms. The non-competitive arms are identified as sub-optimal without having to pull them. We refer to this identification as *implicit* exploration. This implicit exploration is combined with traditional bandit algorithms such as UCB and Thompson sampling to design UCB-C and TS-C algorithms.

3) We perform finite-time regret analysis of our algorithm and show that it performs better than UCB. In fact, the proposed UCB-C algorithm ends up pulling non-competitive arms only \( O(1) \) times. Due to this only \( C−1 \) out of the \( K−1 \) sub-optimal arms are pulled \( O(\log T) \) times, where \( C \) denotes the number of competitive arms. The value of \( C \) can be much less than \( K \) and can even be 1, in which case our algorithm achieves bounded regret! We also prove that our algorithm achieves bounded regret whenever possible.

4) The design of UCB-C makes it easy to extend other classical bandit algorithms (such as Thompson sampling [6], KL-UCB [7], etc.) in the structured bandit setting. The ability to incorporate any classical MAB algorithm is an important advantage of our approach as algorithms such as Thompson sampling have attracted great attention recently due to their superior empirical performance [17]. In particular, the extension to Thompson sampling was deemed to be not immediately possible for the UCB-S algorithm [18] developed for a structured MAB problem similar to ours.

5) We perform experiments on the MOVIELENS dataset to demonstrate the applicability of the UCB-C and TS-C algorithms. Our experimental results show significant improvement over the performance of classical bandit strategies.

**2 Preliminaries**

**2.1 Problem Formulation**

Consider a multi-armed bandit setting with \( K \) arms \( \mathcal{K} = \{1, 2, \ldots, K\} \). At each round \( t \), the player pulls arm \( k_t \in \{1, 2, \ldots, K\} \) and observes the reward \( R_t \). The reward \( R_t \) is a random variable with mean \( \mu_k(\theta) = \mathbb{E}[R_t|\theta, k_t] \), where \( \theta \) is a fixed, but unknown parameter which lies in a known set \( \Theta \), as illustrated in Figure 1. Our formulation allows the set \( \Theta \) to be a countable or uncountable set. The hidden parameter \( \theta \) can also be a vector, however we focus on scalar \( \theta \) for brevity purposes. We denote the unknown true value of parameter \( \theta \) by \( \theta^* \).

The mean reward functions \( \mu_1(\theta), \mu_2(\theta), \ldots, \mu_K(\theta) : \Theta \to \mathbb{R} \) can be arbitrary functions, and they are known to the player. However, the conditional distribution of rewards i.e., \( p(R_t|\theta, k) \) is not known. Throughout the paper, we assume that the rewards \( R_t \) are sub-Gaussian with variance proxy \( \sigma^2 \), i.e., \( \mathbb{E}[\exp(s(R_t - \mathbb{E}[R_t]))] \leq \exp \left( \frac{s^2 \sigma^2}{2} \right) \forall s \in \mathbb{R}, \) and \( \sigma \) is known to the player. Both of these assumptions are common in the multi-armed bandit literature [18][19]. In particular, the sub-Gaussianity of rewards enables us to apply Hoeffding’s inequality, which
Figure 1: Structured bandit setup: mean rewards of different arms share a common hidden parameter $\theta$.

is essential for the analysis of regret (defined below).

The objective of the player is to select arm $k_t$ in round $t$ so as to maximize her cumulative reward $\sum_{t=1}^{T} R_t$ after $T$ slots. If the player had known the true value $\theta^*$, then she would always pull the arm $k^* = \arg\max_{k \in \mathcal{K}} \mu_k(\theta^*)$ which yields the highest mean reward for the parameter $\theta^*$ as that arm would result in the maximum expected cumulative reward. We refer to this arm $k^* = \arg\max_{k \in \mathcal{K}} \mu_k(\theta^*)$ as the optimal arm. Thus, maximizing the cumulative reward is equivalent to minimizing the cumulative regret, which is defined as

$$R_{\text{eg}}(T) \triangleq \sum_{t=1}^{T} \mu_{k^*}(\theta^*) - \mu_{k_t}(\theta^*) = \sum_{k \neq k^*} n_k(T) \Delta_k,$$

where $n_k(T)$ is the number of times arm $k$ is pulled in $T$ slots and $\Delta_k \triangleq \mu_{k^*}(\theta^*) - \mu_k(\theta^*)$ is the sub-optimality gap of arm $k$ or the difference between mean reward of the optimal arm $k^*$ and that of arm $k$. The cumulative regret quantifies the performance of a player in comparison to an oracle that pulls the optimal arm at each round.

Minimizing the cumulative regret is in turn equivalent to minimizing $n_k(T)$, the number of times each sub-optimal arm $k \neq k^*$ is pulled. In this work, we exploit the knowledge of functions $\mu_k(\theta)$ to reduce $n_k(T)$ of certain arms to $O(1)$ instead of the typical $O(\log T)$ scaling.

2.2 Connection with Previously Studied Multi-armed Bandit Frameworks

Since we do not make any assumptions on the functions $\{\mu_1, \mu_2, \ldots, \mu_K\}$, our model subsumes several previously studied frameworks [2][12][13] and is applicable more generally. The similarities and differences between our model and existing works are discussed next.

Classical MAB. Under the classical Multi-armed bandit setting, the rewards obtained from each arm are independent. By considering $\tilde{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ and $\mu_k = \tilde{\theta}_k$, our setting reduces to the classical MAB setting. Our proposed algorithm will in fact perform UCB/Thompson sampling ([1][20]) in this special case.

Structured bandits with linear rewards [13]. In [13], the authors consider a similar model with a common hidden parameter $\theta \in \mathbb{R}$, but the mean reward functions, $\mu_k(\theta)$ are linear in $\theta$. Under this assumption, they design a greedy policy that achieves bounded regret. Our formulation places no such restriction on the reward functions. In the special case when $\mu_k(\theta)$ are linear, our proposed algorithm also achieves bounded regret.

Global Bandits [9]. In [9], a model where mean reward functions are dependent on a common scalar parameter is studied. A key assumption in [9] is that the mean reward functions are invertible and Hölder-continuous. Under these assumptions, they demonstrate that it is possible to achieve bounded regret through a greedy policy. In contrast, our work makes no assumptions on the nature of the functions $\mu_k(\theta)$. In fact, when reward functions are invertible, our algorithms also achieve bounded regret.

Regional Bandits [12]. The paper [12] generalizes global bandits by considering that there are $M$ common unknown parameters, that is, $\theta = (\theta_1, \theta_2, \ldots, \theta_M)$. The mean reward of each arm depends on exactly one of these $M$ parameters, $\theta_m$, and it is also assumed that the mean reward functions are invertible and Hölder-continuous in $\theta_m$. The setting described in [12] is captured in our formulation with vector $\tilde{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)$. Our problem set-up allows arbitrary $\mu_k(\theta)$ that can depend on any subset of the $M$ parameters and need not be invertible or Hölder-continuous.

Finite-armed Generalized Linear bandits [15]. Under the finite-armed linear bandit setting [15], the reward function of arm $x_k$ is $\tilde{\theta}^T x_k$. Here, $\tilde{\theta}$ is the shared unknown parameter. Our proposed algorithms (designed for a general framework) work even in the linear bandit setting and through simulations we demonstrate its performance relative to the GLM-UCB [14] (an algorithm designed for linear bandit setting). For the case when $\mu_k(\theta) = g(\tilde{\theta}^T x_k)$, our setting becomes the generalized linear bandit setting [14], for some known function $g$.

Minimal Exploration in Structured Bandits [21]. The problem formulation in [21] is closely related to this paper. The focus of [21] is to obtain asymptotically optimal regret for the regimes when regret scales as $\log(T)$. When all arms are non-competitive (the bounded regret case), the solution to the optimization problem described in [21] Theorem 1] becomes 0, causing the algorithm to get stuck in the exploitation phase. Also, [21] assume the knowledge of the shape of reward distribution (for example, Gaussian with unknown mean), we only assume that the rewards are sub-Gaussian. Moreover, they assume $\theta \rightarrow \mu_k(\theta)$ is continuous, while we make no such assumption. This also restricts [21] to settings where $\theta$ lies in an uncountable set, whereas our setting allows $\theta$ to lie in a countable set as well.

Finite-armed structured bandits [18]. The work clos-
We illustrate this key idea via the the example shown in Figure 2. In this case, the true parameter \( \theta^* = 3 \), Arm 2 is the optimal arm while Arms 1 and 3 are sub-optimal. An empirical estimate of the mean reward

\[
\hat{\mu}_2(t) = \frac{\sum_{\tau=1}^t R_{\tau \mid k_{\tau} = 2}}{n_2(t)}.
\]

Using this empirical estimate, the player can construct a region in which \( \mu_2(\theta^*) \) lies with high probability. Figure 3 illustrates such a region in shaded pink color. This region can then be used to identify the set of values \( \Theta_t \) within which the true parameter \( \theta^* \) lies with high probability. For example, in Figure 3 that region is the set \([1.5, 4.5]\). If \( \theta^* \) indeed lies in this set, then we can infer that Arm 3 cannot be optimal as its mean reward is less than that of Arm 2 for all \( \theta \in [1.5, 4.5] \). However, Arm 1 may still be better than Arm 2 as it has higher mean reward than Arm 2 for some values of \( \theta \in [1.5, 4.5] \). This provides an example where we implicitly explore Arm 3 without pulling it. As Arm 3 cannot be optimal in the set \([1.5, 4.5]\), we refer to it as non-competitive with respect to the set \([1.5, 4.5]\). On the other hand, we call Arm 1 and 2 competitive with respect to \([1.5, 4.5]\) as they are optimal for at least one \( \theta \) in this set.

We formalize this idea of identifying competitive and non-competitive arms and propose the UCB-C and TS-C algorithms in Section 3 below. In Section 4 we prove that UCB-C can reduce a \( K \)-armed bandit problem to a \( C \)-armed bandit problem, where \( C \leq K \) is the number of competitive arms. In certain regimes (when \( C = 1 \)), we get bounded (i.e., not scaling with \( T \)) regret as we will show in Section 4.

### 3 Proposed Algorithms: UCB-C and TS-C

Classical bandit algorithms such as Thompson sampling and Upper Confidence Bound (UCB) are often termed as index-based policies. At every time instant, these policies maintain an index for each arm, and select the arm with the highest index in the next time slot. More specifically, at each round \( t + 1 \), UCB selects the arm

\[
k_{t+1} = \arg \max_{k \in K} \left( \hat{\mu}_k(t) + \sqrt{\frac{2\alpha \sigma^2 \log(t)}{n_k(t)}} \right),
\]

where \( \hat{\mu}_k(t) \) is the empirical mean of arm \( k \) obtained from the \( n_k(t) \) samples obtained till \( t \).

Under Thompson sampling, we select the arm \( k_{t+1} = \arg \max_{k \in K} S_{k,t} \) at time step \( t \). Here, \( S_{k,t} \) is the sample obtained from the posterior distribution of \( \mu_k \). That is,

\[
k_{t+1} = \arg \max_{k \in K} S_{k,t}, \quad S_{k,t} \sim \mathcal{N} \left( \hat{\mu}_k(t), \frac{\sigma^2}{n_k(t)} \right).
\]

Since mean rewards are correlated through the hidden parameter \( \theta^* \) in the structured bandit model, obtaining an empirical estimate of the mean reward

\[
\hat{\mu}_2(t) = \frac{\sum_{\tau=1}^t R_{\tau \mid k_{\tau} = 2}}{n_2(t)}.
\]
an estimate of $\theta^*$ can help identify the optimal arm. In our approach, we will identify subset of arms, called the competitive arms, through the estimate of $\theta^*$ and then perform UCB or TS over that set of arms. We now define the notions of $\hat{\Theta}$-Competitive and $\tilde{\Theta}$-Non-competitive arms, which are a key component in the design of UCB-C and TS-C Algorithms.

### 3.1 Competitive and Non-Competitive Arms

From the samples observed till time step $t$, one can construct a confidence set $\Theta_t$. The set $\Theta_t$ represents the set of values in which the true parameter $\theta^*$ lies with high confidence, based on rewards observed until time $t$. Next, we define the notions of $\hat{\Theta}$-Competitive and $\tilde{\Theta}$-Non-competitive arms.

**Definition 1 ($\hat{\Theta}$-Competitive arm).** An arm $k$ is said to be $\hat{\Theta}$-Competitive if $\mu_k = \max_{\theta \in \hat{\Theta}} \mu(\theta)$ for some $\theta \in \hat{\Theta}$.

Intuitively, an arm is $\hat{\Theta}$-Competitive if it is optimal for at least one $\theta$ in the confidence set $\hat{\Theta}$. Similarly, we define a $\tilde{\Theta}$-Non-competitive arm as follows.

**Definition 2 ($\tilde{\Theta}$-Non-competitive arm).** An arm $k$ is said to be $\tilde{\Theta}$-Non-competitive if $\mu_k < \max_{\theta \in \tilde{\Theta}} \mu(\theta)$, for all $\theta \in \tilde{\Theta}$.

Intuitively, if an arm is $\tilde{\Theta}$-Non-competitive, it means that it cannot be optimal if the true parameter lies inside the confidence set $\tilde{\Theta}$. This allows us to identify the $\tilde{\Theta}$-Non-competitive arm as sub-optimal under the assumption that the true parameter $\theta^*$ is in the set $\tilde{\Theta}$.

### 3.2 Components of Our Algorithm

Motivated with the above discussion, we propose the following algorithm. At each step $t + 1$, we:

1. Construct a confidence set $\hat{\Theta}_t$ from the samples observed till time step $t$.
2. Identify $\hat{\Theta}_t$-Non-competitive arms.
3. Play a bandit algorithm (UCB or Thompson sampling) among arms which are $\hat{\Theta}_t$-Competitive and choose the next arm $k_{t+1}$ accordingly.

The formal description of this algorithm with UCB and Thompson sampling as final steps is given in Algorithm 1 and Algorithm 2 respectively. Below, we explain the three key components of these algorithms.

**Step 1: Constructing a confidence set, $\hat{\Theta}_t$.** From the samples observed till time step $t$, we define the confidence set as follows:

$$
\hat{\Theta}_t = \left\{ \theta : \forall k \in \mathcal{K}, |\mu_k(\theta) - \hat{\mu}_k| < \sqrt{\frac{2\alpha \sigma^2 \log t}{n_k(t)}} \right\}.
$$

### Algorithm 1 UCB-C Correlated UCB Algorithm

1: **Input:** Reward Functions $\{\mu_1, \mu_2, \ldots, \mu_K\}$
2: **Initialize:** $n_k = 0$ for all $k \in \{1, 2, \ldots, K\}$
3: for each round $t + 1$ do
4:  **Confidence set construction:**
5:    $$
\hat{\Theta}_t = \left\{ \theta : \forall k \in \mathcal{K}, |\mu_k(\theta) - \hat{\mu}_k(t)| < \sqrt{\frac{2\alpha \sigma^2 \log t}{n_k(t)}} \right\}.
$$
6:  **Define competitive set $C_t$:**
7:    $$
C_t = \left\{ k : \mu_k(\theta) = \max_{\theta \in \hat{\Theta}_t} \mu(\theta) \text{ for some } \theta \in \hat{\Theta}_t \right\}.
$$
8:  **UCB among competitive arms**
9:    $$
k_{t+1} = \arg \max_{k \in C_t} \left( \hat{\mu}_k(t) + \sqrt{\frac{2\alpha \sigma^2 \log t}{n_k(t)}} \right).
$$
10: **Update empirical mean** $\hat{\mu}_k(t+1)$ and $n_k(t+1)$ for every arm $k$.
11: end for

Here $\hat{\mu}_k$ is the empirical mean of rewards obtained in $n_k$ samples of arm $k$. Using the confidence bound on mean of each arm, we construct $\hat{\Theta}_t$ that contains values of $\theta$, which lead to mean reward within confidence interval of mean of each arm.

**Step 2: Identifying $\hat{\Theta}_t$-Non-competitive arms.** At each time step $t + 1$, we define the set $C_t$ as the set of $\hat{\Theta}_t$-Competitive arms, that includes all arms $k$ that satisfy $\mu_k = \max_{\theta \in \hat{\Theta}_t} \mu(\theta)$ for some $\theta \in \hat{\Theta}_t$. The rest of the arms, termed as $\tilde{\Theta}_t$-Non-competitive, are eliminated for round $t + 1$ and are not considered in the next part of the algorithm. Note that these arms can be $\hat{\Theta}_t$-Competitive in subsequent rounds.

**Step 3: Play bandit algorithm among $\hat{\Theta}_t$-Competitive arms.** After identifying the $\hat{\Theta}_t$-Competitive arms, we use classical bandit algorithms such as UCB and Thompson sampling to decide which arm to play at time step $t + 1$ as specified in (2) and (3) respectively.

It is important to note that the last step of our algorithm can utilize any one of the classical bandit algorithms. This allows us to easily define a Thompson sampling algorithm which has attracted great attention [6, 17, 22] for the structured bandits problem considered in this paper. The ability to employ any bandit algorithm in its last step is an important advantage of our algorithm. For instance, the extension to Thompson sampling was deemed to be not possible for the UCB-S algorithm proposed in [18].
Algorithm 2 TS-C Correlated Thompson sampling

1: Steps 1 to 5 as in Algorithm 1
2: Apply Thompson sampling on $C_t$
3: for $k \in C_t$ do
4: Sample $S_{k,t} \sim N\left(\hat{\mu}_k(t), \frac{\sigma^2}{n_k(t)}\right)$.
5: end for
6: $k_{t+1} = \arg\max_{k \in C_t} S_{k,t}$
7: Update empirical mean $\hat{\mu}_k$ and $n_k$ for all arm $k$.

4 Regret Analysis and Bounds

In this section, we evaluate the performance of the UCB-C algorithm through a finite-time analysis of the expected cumulative regret defined as

$$
\mathbb{E}\left[\text{Reg}(T)\right] = \sum_{k=1}^{K} \mathbb{E}\left[n_k(T)\right] \Delta_k,
$$

where $\Delta_k = \mu_k^*(\theta^*) - \mu_k(\theta^*)$ and $n_k(T)$ is the number of times arm $k$ is pulled in total of $T$ time steps. To analyze the expected regret, we need to determine $\mathbb{E}\left[n_k(T)\right]$ for each sub-optimal arm $k \neq k^*$.

In this section we derive $\mathbb{E}\left[n_k(T)\right]$ separately for competitive and non-competitive arms and show that it is $O(1)$ for non-competitive arms. For the purpose of regret analysis, we first define confidence set $\Theta^*$ and formally define the notion of competitive and non-competitive arms. Define the set $\Theta^* = \{\theta: \mu_{k^*}(\theta) = \mu_{k^*}(\theta^*)\}$. We can view $\Theta^*$ as the confidence set $\Theta_t$ after the optimal arm is sampled infinitely many times. A $\Theta^{*\epsilon}(\cdot)$-non-competitive arm is then defined as follows.

Definition 3 ($\Theta^{\epsilon\epsilon}(\cdot)$-non-competitive arm). We call an arm $k$ as $\Theta^{\epsilon\epsilon}(\cdot)$-non-competitive if

$$
\mu_{k^*}(\theta) > \mu_k(\theta), \text{ for all } \theta: |\mu_{k^*}(\theta^*) - \mu_{k^*}(\theta)| < \epsilon,
$$

where $\Theta^{\epsilon\epsilon}(\cdot)$ is an expanded version of $\Theta^*$ that allows a deviation of at most $\epsilon$ from the optimal mean reward. For each arm $k$, we define $\epsilon_k$ as the largest $\epsilon$ for which it is $\Theta^{\epsilon\epsilon}(\cdot)$-non-competitive, where the set $\Theta^{\epsilon\epsilon}(\cdot)$ is the expanded version of $\Theta^*$. An arm is competitive (according to Definition 1) if $\epsilon_k = 0$, otherwise it is said to be $\Theta^{\epsilon\epsilon}(\cdot)$-Non-Competitive.

Our first result shows that expected pulls for any arm (competitive or non-competitive) is $O(\log T)$.

Theorem 1 (Expected pulls for any arm). The expected number of times any arm is pulled by UCB-C Algorithm is upper bounded as

$$
\mathbb{E}\left[n_k(T)\right] \leq 8\alpha\sigma^2 \frac{\log T}{T^2} + \frac{2\alpha}{\alpha - 2} + \sum_{t=1}^{T} 2Kt^{1-\alpha}
$$

$$
= O(\log T) \text{ for } \alpha > 2,
$$

where

$$
t_0 = \inf \left\{ t \geq 2: \Delta_k \geq 4 \frac{K\alpha \sigma^2 \log \tau}{\tau}, \epsilon_k \geq \sqrt{8 \alpha \sigma^2 K \log \tau} \right\}.
$$

Plugging the results of Theorem 1 and Theorem 2 in (5) yields the following bound on the expected regret of the UCB-C Algorithm. Note that in this work we consider a finite-armed setting and hence are focused on studying the scaling of regret with respect to $T$ and not with $K$.

Theorem 2 (Expected pulls of Non-competitive Arms). If an arm $k$ is $\Theta^{\epsilon\epsilon}(\cdot)$-non-competitive for some $\epsilon_k > 0$, then the number of times it is pulled by UCB-C is upper bounded as

$$
\mathbb{E}\left[n_k(T)\right] \leq Kt_0 + \sum_{t=1}^{T} 2Kt^{1-\alpha} + K^3 \sum_{t=Kt_0}^{T} 6 \left(\frac{t}{K}\right)^{2-\alpha}
$$

$$
= O(1) \text{ for } \alpha > 3,
$$

where

$$
t_0 = \inf \left\{ \tau \geq 2: \Delta_k \geq 4 \frac{K\alpha \sigma^2 \log \tau}{\tau}, \epsilon_k \geq \sqrt{8 \alpha \sigma^2 K \log \tau} \right\}.
$$

Our next result shows that the expected number of pulls for an $\Theta^{\epsilon\epsilon}(\cdot)$-non-competitive arm are bounded.

Theorem 3 (Regret upper bound). For $\alpha > 3$, the expected regret of the UCB-C Algorithm is upper bounded as

$$
\mathbb{E}\left[\text{Reg}(T)\right] \leq \sum_{k \in C\setminus\{k^*\}} \Delta_k t_k^{(c)}(T) + \sum_{k \in C\setminus\{k^*\}} \Delta_k U_k^{(nc)}(T),
$$

$$
= (C - 1) \cdot O(\log T) + O(1),
$$

Here, $U_k^{(c)}(T)$ is the upper bound on $\mathbb{E}\left[n_k(T)\right]$ given in Theorem 1 and $U_k^{(nc)}(T)$ is the upper bound on $\mathbb{E}\left[n_k(T)\right]$ given in Theorem 2. The set $C$ denotes the set of competitive arms and $C$ is the cardinality of that set.

Reduction in the effective number of arms. The classic UCB algorithm that is agnostic to the structure of the problem pulls each of the $(K - 1)$ sub-optimal arms $O(\log T)$ times. In contrast, our algorithm pulls only $(C - 1)$ sub-optimal arms $O(\log T)$ times, where $C \leq K$. In fact, when $C = 1$, all sub-optimal arms are pulled only $O(1)$ times, leading to a bounded regret. The value of $C$ depends on the specific correlation structure, i.e.,
the functions \( \mu_k(\theta) \), and the hidden parameter \( \theta^* \). Cases with \( C = 1 \) can arise quite often in practical settings. For example, when the optimal arm \( k^* \) is invertible around \( \theta^* \), the set \( \Theta^* \) becomes a singleton; i.e., there is just a single \( \theta \in \Theta \) that leads to \( \mu_k, (\theta^*) \). In that case, all sub-optimal arms become non-competitive and our UCB-C algorithm returns bounded (i.e., \( O(1) \)) regret.

We now show that the UCB-C algorithm achieves bounded regret whenever possible. We do so by analyzing a lower bound obtained in [21].

**Proposition 1 (Lower bound).** For any uniformly good algorithm \( \Pi \), and for any \( \theta \in \Theta \), we have:

\[
\lim_{T \to \infty} \inf \frac{\text{Reg}(T)}{\log T} \geq L(\theta), \quad \text{where}
\]

\[
L(\theta) = \begin{cases} 
0 & \text{if } C = 1, \\
> 0 & \text{if } C > 1.
\end{cases}
\]

An algorithm \( \pi \) is uniformly good if \( \text{Reg}^\pi(T, \theta) = o(T^a) \) for all \( a > 0 \) and all \( \theta \in \Theta \).

The proof of this proposition, given in Appendix, follows from a lower bound derived in [21]. This lower bound leads us to the following observation.

**Remark 1 (Bounded regret whenever possible).** The result on lower bound in Proposition 7 shows that sub-logarithmic regret is possible only when \( C = 1 \). In the case when \( C = 1 \), our proposed algorithm achieves a bounded regret (see Theorem 3). This implies that the UCB-C algorithm is able to achieve bounded regret whenever possible and reduce the number of effective arms for the cases when regret is logarithmic.

## 5 Simulation Results

We now study the empirical performance of the proposed algorithms. For all simulations we choose \( \alpha = 3 \). Rewards are drawn from the distribution \( \mathcal{N}(\mu_k(\theta^*), 4) \), i.e., \( \sigma = 2 \). For each case considered, we run 100 independent experiments and present the average regret. We demonstrate i) how UCB-C reduces the number of arms selected \( O(\log T) \) times; ii) the comparison with UCB-S [18]; iii) how our algorithms perform for multi-dimensional \( \theta \); and iv) the performance of our algorithms in the linear bandit setting.

**Reduction in the effective number of arms.** In Figure 4 we compare the regret of the UCB-C algorithm with classic UCB for the example considered in Figure 3. For \( \theta^* = 0.5 \), Arm 2 is optimal and Arms 1 and 3 are non-competitive. As expected from our regret analysis, in Figure 5(a) the UCB-C algorithm achieves bounded regret, while the regret of UCB grows logarithmically in the number of time steps \( T \). When \( \theta^* = 1.8 \), Arm 3 is optimal, Arm 2 becomes competitive and Arm 1 is non-competitive. In this case, we expect UCB-C to pull Arm 1 only \( O(1) \) times due to which we notice significantly reduced regret with UCB-C as compared to UCB in Figure 5(b). Figure 5(c) shows the case where \( \theta^* = 2.8 \), leading to Arm 1 being optimal and all the arms being competitive. As the UCB-C algorithm uses samples from all arms to generate the confidence set \( \hat{\Theta}_t \), it achieves empirically smaller regret for this setting than the UCB algorithm even when all arms are competitive.

**Comparison with UCB-S [18].** We now compare the performance of our UCB-C and TS-C algorithms against the UCB and UCB-S Algorithm proposed in [18]. We consider the example shown in Figure 4. We plot the cumulative regret of the UCB, UCB-S, UCB-C and TS-C algorithms over 50000 time steps for the values of \( \theta^* \) between 0 and 5 in Figure 7. When \( \theta^* \) is below 3, UCB-S, UCB-C and TS-C all obtain smaller regret than UCB as they are able to identify the sub-optimal arm as non-competitive. For \( \theta^* \in (3, 5) \), UCB-C has a performance similar to that of UCB as sub-optimal arm is competitive.

We see that when \( \theta^* = 3.25 \), UCB-S achieves a regret which is quite large compared to even UCB. This is because with \( \theta^* = 3.25 \), Arm 1 is optimal but the samples from Arm 1 does not allow accurate estimation of \( \theta \) as its mean reward is constant over a large region of \( \theta \) values.
Figure 6: Arm 2 is optimal for $\theta^* \in [0, 3]$ and Arm 1 is optimal for $\theta^* \in [3, 5]$.

Figure 7: Cumulative regret of UCB, UCB-S, UCB-C and TS-C for the example in Figure 6 over 50000 runs compared over different values of $\theta^*$.

(at some of which Arm 2 is optimal). Since the UCB-S algorithm selects $k_{t+1} = \arg \max_{k \in K} \sup_{\theta \in \Theta_k} \mu_k(\theta)$, the sup leads it to select the sub-optimal Arm 2 a large number of times. We also see that TS-C does not suffer from the same issue and performs significantly better than all other algorithms. We attribute this to the fact that Thompson sampling can offer significant empirical improvements over UCB. This highlights the benefits of being able to incorporate Thompson sampling in a correlated MAB algorithm which was deemed to be not possible in [18].

Multi-Dimensional $\theta$. Our formulation allows for $\theta$ to be a vector as well. To demonstrate this, we consider a case where $\vec{\theta} = (\theta_1, \theta_2)$ and $\theta_1, \theta_2 \in [-1, 1]$. The three arms have mean reward functions $\mu_1(\vec{\theta}) = \theta_1 + \theta_2$, $\mu_2(\vec{\theta}) = \theta_1 - \theta_2$ and $\mu_3(\vec{\theta}) = \max(|\theta_1|, |\theta_2|)$. Note that this setting cannot be captured by linear bandit or generalized linear bandit models. Figure 8 shows the performance of UCB-C and TS-C algorithms against the UCB algorithm in this setting. For $\vec{\theta}^* = (0.9, 0.2)$, Arm 1 is optimal and Arm 3 is Non-Competitive, due to which we see empirical superiority of UCB-C and TS-C algorithms over UCB. For the case where $\vec{\theta}^* = (-0.2, 0.1)$, Arm 3 is optimal and all arms are competitive. As as result, UCB-C and UCB give the same performance, while TS-C performs better due to the empirical superiority of the Thompson sampling in this setting.

Comparison with Finite-Armed Linear Bandits. Under the finite-armed linear bandit setting, the reward function of arm $k$ is $\vec{\theta}^* x_k$, where $x_k$ is a known vector associated with arm $k$. When $\mu_k(\vec{\theta}) = \vec{\theta}^* x_k$, our setting covers the linear bandit setup. We now compare the performance of our UCB-C and TS-C algorithms against the GLM-UCB algorithm, which is specifically designed for linear bandits. In the example considered, $\Theta = [0, 1] \times [0, 1]$ and we consider three arms $x_1 = (2, 1), x_2 = (1, 1.5)$ and $x_3 = (3, -1)$. Figure 9 shows the results of the simulation performed. We see that when hidden parameter $\vec{\theta} = (0.9, 0.9)$, our algorithm outperforms the GLM-UCB algorithm [14]. In the case when $\vec{\theta} = (0.5, 0.5)$, GLM-UCB has better performance than UCB-C however TS-C performs the best. This occurs due to the fact that Thompson sampling is empirically better than UCB in this setting. It is important to note that while UCB-C and TS-C are designed for a much broader class of problems, they still show competitive performance relative to specialized algorithms (i.e., GLM-UCB) in the linear bandit setting.
6 Experiment with MOVIELENS dataset

We now show the performance of UCB-C and TS-C on a real-world dataset. As mentioned earlier, structured bandits might be useful in targeted advertising without observing the user context. We use the MOVIELENS dataset [23], which contains a total of 1M ratings made by 6040 users for 3883 movies to demonstrate how UCB-C and TS-C can be deployed in practice. In particular, we will demonstrate the superiority of our algorithms over the classical UCB in movie recommendations.

We consider the MOVIELENS dataset, in which there are 106 different types of users (based on their age and occupation features) and there are 18 different genres of movies. We refer to each type of user as a meta-user (so, there are 106 meta-users), which corresponds to the parameter $\theta$ in our setup. For example, one meta-user in the data-set contains 18-25 year old college students. The users have given ratings to the movies on a scale of 1 to 5. Each movie is associated with one (and in some cases, multiple) genres. For the experiments, of the possibly multiple genres for each movie one is randomly allotted.

For a particular user whose features is unknown (i.e., the true value of the $\theta$ is hidden), our goal is to recommend a movie sequentially from the genre (i.e., arm) that the user has the highest mean rating for. We use a part of the dataset (50%) as the training dataset, on which the mean reward mappings from meta-users to different genres are learned. The learned mappings from meta users ($\theta$) to genres (arms) are shown in the appendix. The learned mappings indicate that the mean-reward mappings of meta-users for different genres are related to one another. For example, on average 56+ year old retired users tend to like documentaries and tend to dislike children’s movies. Such dependencies are learned in the mapping from $\theta$ to $\mu_k(\theta)$ in training. In order to implement it for recommendations or advertising, these mappings could be learned from surveys in which users participate with their consent.

The goal is to recommend the most preferred genre to an unknown meta-user. To test the algorithm, we select a meta-user $\theta^*$ from the remaining 50% of the data, i.e., the test data. At every round $t$, the MAB algorithm selects a genre to recommend to the meta-user. The rating corresponding to that genre is obtained by sampling randomly from the available reward samples of the (meta-user, genre) pairing. This process is repeated for a total of 15000 rounds. We measure the regret of algorithm in the test data by evaluating $\sum_{t=1}^{T} \mu_{k^*}(\theta^*) - \mu_{k^*}(\theta^*)$ on the test data. We report the results in Figure 10 for two different types of meta-users (i.e., $\theta^*$), i) 18-25 year old college students and ii) 25-34 year old executives.

Figure 10 shows that UCB-C and TS-C algorithms are able to achieve significantly lower regret than UCB as only a few arms are pulled $O(\log T)$ times. We plot results for two different values of $\theta^*$, similar improvements are seen for other values of $\theta^*$ as well. This experiment demonstrates that it is indeed possible to provide better recommendations than UCB without viewing the contextual (and private) information of the users. Our approach can be useful in broader settings for advertising or recommendations without using personal information.

7 Concluding Remarks

In this work, we studied a structured bandit problem in which the mean rewards of different arms are related through a common hidden parameter. This is done under a general framework without any assumptions on the mean reward functions, due to which our model subsumes several previously studied frameworks [9, 12, 13].

We developed an approach that allows identifying some arms as sub-optimal without exploring them explicitly. Through finite time regret analysis, we showed that the resulting UCB-C algorithm reduces the $K$-armed bandit problem to a $C$-armed bandit problem in the sense that only $C - 1 \leq K - 1$ of the sub-optimal arms are chosen $O(\log T)$ times. We also showed that UCB-C achieves bounded regret whenever possible. Our approach allows extension of any classical bandit strategy in the structured bandit setting. Through experiments on the MOVIELENS dataset, we demonstrated the empirical superiority of UCB-C and TS-C algorithms over other approaches.

Ongoing work includes the finite-time regret analysis of TS-C algorithm. An interesting direction would be to study the case where $\theta$ is random with an unknown distribution. We also plan to study the best-arm identification problem in the same setting.
References

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

[2] Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[3] Cem Tekin and Eralp Turgay. Multi-objective contextual multi-armed bandit problem with a dominant objective. *arXiv preprint arXiv:1708.05655*, 2017.

[4] J. Nino-Mora. *Stochastic scheduling*. In *Encyclopedia of Optimization*, pages 3818–3824. Springer, New York, 2 edition, 2009.

[5] John White. *Bandit algorithms for website optimization*. “O’Reilly Media, Inc.”, 2012.

[6] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[7] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

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

[9] Onur Atan, Cem Tekin, and Mihaela van der Schaar. Global multi-armed bandits with Hölder continuity. In *AISTATS*, 2015.

[10] Jian Huang, Mingming Leng, and Mahmut Parlar. Demand functions in decision modeling: A comprehensive survey and research directions. *Decision Sciences*, 44(3):557–609, 2013.

[11] Cong Shen, Ruida Zhou, Cem Tekin, and Mihaela van der Schaar. Generalized global bandit and its application in cellular coverage optimization. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):218–232, 2018.

[12] Zhiyang Wang, Ruida Zhou, and Cong Shen. Regional multi-armed bandits. In *AISTATS*, 2018.

[13] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis. A structured multi-armed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802, Dec 2009.

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

[15] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. *arXiv preprint arXiv:1610.04491*, 2016.

[16] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[17] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

[18] Tor Lattimore and Rémi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558, 2014.

[19] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil’ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

[20] W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3-4):285–294, December 1933.

[21] Richard Combes, Stefan Magureanu, and Alexandre Proutiè re. Minimal exploration in structured stochastic bandits. In *NIPS*, 2017.

[22] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *CoRR*, abs/1707.02038, 2017.

[23] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5, 4, Article 19, 2015.

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

A Lower bound

We use the following result of [21] to state Proposition 1.

Theorem 4 (Lower bound, Theorem 1 in [21]). For any uniformly good algorithm [1], and for any \( \theta \in \Theta \), we have:

\[
\lim \inf_{T \to \infty} \frac{\text{Reg}(T)}{\log T} \geq L(\theta),
\]

where \( L(\theta) \) is the solution of the optimization problem:

\[
\min_{D(\theta, \lambda, k) \geq 0, k \in K} \sum_{k \in K} \eta(k) \left( \max_{\ell \in K} \mu_{\ell}(\theta) - \mu_k(\theta) \right)
\]

subject to \( \sum_{k \in K} \eta(k) D(\theta, \lambda, k) \geq 1, \forall \lambda \in \Lambda(\Theta), \)

\[
\Lambda(\theta) = \{ \lambda \in \Theta^*: k^* \neq \arg \max_{k \in K} \mu_k(\lambda) \}. \tag{10}
\]

Here, \( D(\theta, \lambda, k) \) is the KL-Divergence between distributions \( f_R(R_{k, t}(\theta, k)) \) and \( f_R(R_{k, t}(\lambda, k)) \). An algorithm, \( \pi \), is uniformly good if \( \text{Reg}^\pi(T, \theta) = o(T^a) \) for all \( a > 0 \) and all \( \theta \in \Theta \).

We see that the solution to the optimization problem (10) is \( L(\theta) = 0 \) only when the set \( \Lambda(\theta) \) is empty. The set \( \Lambda(\theta) \) being empty corresponds to a case where all sub-optimal arms are non-competitive. This implies that sub-logarithmic regret is possible only when \( C = 1 \), i.e., there is only one competitive arm, which is the optimal arm, and all other arms are non-competitive. It is assumed that reward distribution of an arm \( k \) is parameterized by the mean \( \mu_k \) of arm \( k \); this ensures that if \( \mu_k(\theta) = \mu_k(\lambda) \) then we have \( D(\theta, \lambda, k) = 0 \).

B Proof for the UCB-C Algorithm

Fact 1 (Hoeffding’s inequality). Let \( Z_1, Z_2, \ldots, Z_T \) be i.i.d. random variables, where \( Z_i \) is \( \sigma^2 \) sub-gaussian with mean \( \mu \), then

\[
\Pr( | \hat{\mu} - \mu | \geq \epsilon ) \leq 2 \exp \left( -\frac{\epsilon^2 T}{2\sigma^2} \right).
\]

Here \( \hat{\mu} \) is the empirical mean of the \( Z_1, Z_2, \ldots, Z_T \).

Lemma 1 (Standard result used in bandit literature). If \( \hat{\mu}_{k,n_k(t)} \) denotes the empirical mean of arm \( k \) by pulling arm \( k \) \( n_k(t) \) times through any algorithm and \( \mu_k \) denotes the mean reward of arm \( k \), then we have

\[
\Pr( | \hat{\mu}_{k,n_k(t)} - \mu_k | \geq \epsilon, \tau_2 \geq n_k(t) \geq \tau_1 ) \leq 2 \exp \left( -\frac{\epsilon^2}{2\sigma^2} \right) \tag{12}
\]

Proof. Let \( Z_1, Z_2, \ldots, Z_t \) be the reward samples of arm \( k \) drawn separately. If the algorithm chooses to play arm \( k \) for \( m \)th time, then it observes reward \( Z_m \). Then the probability of observing the event \( \{ \hat{\mu}_{k,n_k(t)} - \mu_k \geq \epsilon, \tau_2 \geq n_k(t) \geq \tau_1 \} \) can be upper bounded as follows,

\[
\Pr( \hat{\mu}_{k,n_k(t)} - \mu_k \geq \epsilon, \tau_2 \geq n_k(t) \geq \tau_1 ) = \Pr \left( \left( \hat{\mu}_{k,n_k(t)} - \mu_k \geq \epsilon \right), \tau_2 \geq n_k(t) \geq \tau_1 \right)
\]

\[
\leq \Pr \left( \left( \bigcup_{m=\tau_1}^{\tau_2} \left\{ \sum_{i=1}^{m} \frac{Z_i}{m} - \mu_k \geq \epsilon \right\} \right), \tau_2 \geq n_k(t) \geq \tau_1 \right)
\]

\[
\leq \sum_{s=\tau_1}^{\tau_2} \exp \left( -\frac{s \epsilon^2}{2n_k(t)} \right). \tag{18}
\]

Lemma 2. The Probability that the difference between the true mean of arm \( k \) and its empirical mean after \( t \) time slots is more than \( \sqrt{\frac{2\alpha \sigma^2 \log t}{n_k(t)}} \) is upper bounded by \( 2t^{1-\alpha} \), i.e.,

\[
\Pr \left( | \mu_k(\theta^*) - \hat{\mu}_k | \geq \sqrt{\frac{2\alpha \sigma^2 \log t}{n_k(t)}} \right) \leq 2t^{1-\alpha}.
\]

Proof. See that,

\[
\Pr \left( | \mu_k(\theta^*) - \hat{\mu}_{k,n_k(t)} | \geq \sqrt{\frac{2\alpha \sigma^2 \log t}{n_k(t)}} \right) \leq \sum_{m=1}^{t} \Pr \left( | \mu_k(\theta^*) - \hat{\mu}_{k,m} | \geq \sqrt{\frac{2\alpha \sigma^2 \log t}{m}} \right) \tag{19}
\]

\[
\leq \sum_{m=1}^{t} 2t^{-\alpha} \tag{20}
\]

\[
= 2t^{1-\alpha}. \tag{21}
\]
We have (19) from union bound and is a standard trick to deal with the random variable \( n_k(t) \) as it can take values from 1 to \( t \) (Lemma 1). The true mean of arm \( k \) is \( \mu_k(\theta^*) \). Therefore, if \( \hat{\mu}_{k,m} \) denotes the empirical mean of arm \( k \) taken over \( m \) pulls of arm \( k \) then, (20) follows from Fact 1 with \( \epsilon \) in Fact 1 being equal to \( \sqrt{\frac{2\alpha^2 \log t}{n_k(t)}} \).

\[
\text{Lemma 3. Define } E_1(t) \text{ to be the event that arm } k^* \text{ is } \Theta_t\text{-non-competitive for the round } t+1, \text{ then, }
\Pr(E_1(t)) \leq 2Kt^{1-\alpha}.
\]

\[
\text{Proof. Observe that,}
\begin{align*}
\Pr(E_1(t)) &\leq \Pr(\theta^* \notin \hat{\Theta}_t) \\
&= \Pr(\bigcup_{k \in K} |\mu_k(\theta^*) - \hat{\mu}_{k,n_k(t)}| \geq \sqrt{\frac{2\alpha^2 \log t}{n_k(t)}}) \\
&\leq \sum_{k=1}^{K} \Pr(|\mu_k(\theta^*) - \hat{\mu}_{k,n_k(t)}| \geq \sqrt{\frac{2\alpha^2 \log t}{n_k(t)}}) \\
&\leq \sum_{k=1}^{K} \sum_{m=1}^{t} \Pr(|\mu_k(\theta^*) - \hat{\mu}_{k,m}| \geq \sqrt{\frac{2\alpha^2 \log t}{m}}) \\
&\leq K \sum_{m=1}^{t} 2t^{-\alpha} \\
&= 2Kt^{1-\alpha}.
\end{align*}
\]

We are using \( \hat{\mu}_{k,m} \) to denote the empirical mean of rewards from arm \( k \) obtained from it’s \( m \) pulls. Here (23) follows from definition of confidence set and (24) follows from union bound. We have (25) from union bound and is a standard trick to deal with the random variable \( n_k(t) \) as it can take values from 1 to \( t \) (Lemma 1). Inequality (26) follows from Hoeffding’s lemma.

\[
\text{Lemma 4. If } \Delta_{\min} \geq \sqrt{\frac{K\alpha^2 \log t}{t_0}} \text{ for some constant } t_0 > 0, \text{ then,}
\Pr(k_{t+1} = k, n_k(t) \geq s) \leq (2K+4)t^{1-\alpha} \text{ for } s \geq \frac{t_0}{2K},
\forall t > t_0, \text{ where } k \neq k^* \text{ is a suboptimal arm.}
\]

\[
\text{Proof. The probability that arm } k \text{ is pulled at step } t + 1, \text{ given it has been pulled } s \text{ times can be bounded as follows:}
\begin{align*}
\Pr(k_{t+1} = k, n_k(t) \geq s) &= \Pr(I_k(t) = \max_{k' \in C} I_{k'}(t), n_k(t) \geq s) \\
&\leq \Pr(E_1(t)) \cup (E_1^c(t), I_k(t) > I_{k^*}(t), n_k(t) \geq s) \\
&\leq \Pr(E_1(t), n_k(t) \geq s) + \Pr(E_1^c(t), I_k(t) > I_{k^*}(t), n_k(t) \geq s) \\
&\leq 2Kt^{1-\alpha} + \Pr(I_k(t) > I_{k^*}(t), n_k(t) \geq s) \\
&\leq \Pr(I_k(t) > I_{k^*}(t), n_k(t) \geq s) + 2t^{1-\alpha} \\
&\leq \Pr(\hat{\mu}_k + \sqrt{\frac{2\alpha^2 \log t}{n_k(t)}} > \mu_{k^*}, n_k(t) \geq s) + 2t^{1-\alpha} \\
&= \Pr(\hat{\mu}_k - \mu_k(\theta^*) > \Delta_k - \sqrt{\frac{2\alpha^2 \log t}{n_k(t)}}, n_k(t) \geq s) + 2t^{1-\alpha} \\
&\leq 2t \exp \left(-\frac{s}{\alpha^2} \left(\Delta_k - \sqrt{\frac{2\alpha^2 \log t}{s}}\right)^2\right) + 2t^{1-\alpha} \\
&= 2t^{1-\alpha} \exp \left(-\frac{s}{2\alpha^2} \left(\Delta_k^2 - 2\Delta_k \sqrt{\frac{2\alpha^2 \log t}{s}}\right)\right) + 2t^{1-\alpha} \\
&= 4t^{1-\alpha} \text{ for all } t > t_0.
\end{align*}
\]

Equation (32) follows from the fact that \( P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) \). Inequality (33) arises from dropping \( P(B) \) and \( P(A|B^c) \) in the previous expression. We have (34) from Lemma 3 and the fact that \( I_k(t) = \hat{\mu}_k + \sqrt{\frac{2\alpha^2 \log t}{n_k(t)}} \). Inequality (38) follows from the Hoeffding’s inequality and the term \( t \) before the exponent in (38) arises as the random variable \( n_k(t) \), can take
values between $s$ and $t$ (Lemma 4). Equation (40) results from the definition of $t_0$ and the fact that $s > \frac{1}{2\kappa}$.

Plugging the result of (40) in the expression (31) completes the proof of Lemma 4.

Lemma 5. Consider a suboptimal arm $k \neq k^*$, which is $\Theta^*(\epsilon_k)$-non-competitive. If $\epsilon_k \geq \sqrt{\frac{8\alpha^2 \kappa \log t_0}{t_0}}$ for some constant $t_0 > 0$, then,

$$\Pr(k_{t+1} = k, k^* = k^{\text{max}}) \leq 2t^{1-\alpha},$$

where $k^{\text{max}} = \arg \max_{k \in K} n_k(t)$.

Proof. We now bound this probability as,

$$\Pr(k_{t+1} = k, k^* = k^{\text{max}}) = \Pr(k \in C_t, I_k = \max_{I_k} k^*, k^* = k^{\text{max}})$$

$$\leq \Pr(k \in C_t, k^* = k^{\text{max}})$$

$$\leq \Pr( |\mu_k - \mu_{k^*}(\theta^*)| > \epsilon_k, k^* = k^{\text{max}})$$

$$\leq 2t \exp \left( -\frac{\epsilon^2 t}{8K\sigma^2} \right)$$

$$\leq 2t^{1-\alpha} \quad \forall t > t_0.$$ (45)

See that $|\mu_k - \mu_{k^{\text{max}}}| < \frac{\epsilon_k}{2} \Rightarrow |\mu_k - \mu_k(\theta^*)| < \epsilon$ for $\theta \in \Theta_t$. This holds as $\sqrt{\frac{2\alpha^2 \log t_0}{t_0}} \leq \frac{\epsilon_k}{2}$ and if $\theta \in \Theta_t$, then $|\mu_k - \mu_k(\theta^*)| \leq \sqrt{\frac{2\alpha^2 \log t_0}{t_0}} \leq \frac{\epsilon_k}{2}$. Therefore in order for arm $k$ to be $\Theta_t$-competitive, we need at least $|\mu_k - \mu_k(\theta^*)| > \epsilon_k/2$, which leads to (43) as arm $k$ is $\epsilon_k$ non-competitive. Inequality (44) follows from Hoeffding’s inequality. The term $t$ before the exponent in (43) arises as the random variable $n_k^{\tau}$ can take values from $\frac{1}{2}$ to $t$ (Lemma 4).

Lemma 6. Let $t_0$ be the minimum integer satisfying $\Delta_{\text{min}} \geq 4\sqrt{\frac{K\alpha^2 \log t_0}{t_0}}$ then $\forall t > Kt_0$, and $\forall k \neq k^*$, we have,

$$\Pr\left( n_k(t) > \frac{t}{K} \right) \leq 6K^2 \left( \frac{t}{K} \right)^{2-\alpha}.$$ (46)

Proof. We expand $\Pr\left( n_k(t) > \frac{t}{K} \right)$ as,

$$\Pr\left( n_k(t) > \frac{t}{K} \right) = \left( \Pr\left( n_k(t) > \frac{t}{K} \mid n_k(t-1) \geq \frac{t}{K} \right) \right) +$$

$$\left( \Pr\left( k_t = k \mid n_k(t-1) = \frac{t}{K} \right) \right) \times$$

$$\left( \Pr\left( n_k(t-1) = \frac{n}{K} - 1 \right) \right)$$

Here (48) follows from Lemma 4.

This gives us $\forall (t-1) > t_0$, we have,

$$\Pr\left( n_k(t) > \frac{t}{K} \right) - \Pr\left( n_k(t-1) \geq \frac{t}{K} \right) \leq 6K(t-1)^{1-\alpha}.$$ (49)

Now consider the summation

$$\sum_{\tau = \frac{t}{K}}^{t} \Pr\left( n_k(\tau) \geq \frac{t}{K} \right) - \Pr\left( n_k(\tau-1) \geq \frac{t}{K} \right)$$

$$\leq \sum_{\tau = \frac{t}{K}}^{t} 6K(\tau-1)^{1-\alpha}.$$ (50)

This gives us,

$$\Pr\left( n_k(t) \geq \frac{t}{K} \right) - \Pr\left( n_k\left( \frac{t}{K} - 1 \right) \geq \frac{t}{K} \right)$$

$$\leq \sum_{\tau = \frac{t}{K}}^{t} 6K(\tau-1)^{1-\alpha}.$$ (51)

Since $\Pr\left( n_k\left( \frac{t}{K} - 1 \right) \geq \frac{t}{K} \right) = 0$, we have,

$$\Pr\left( n_k(t) \geq \frac{t}{K} \right) \leq \sum_{\tau = \frac{t}{K}}^{t} 6K(\tau-1)^{1-\alpha} \leq 6K^2 \left( \frac{t}{K} \right)^{2-\alpha}.$$ (52)
**Proof of Theorem 2** We bound \( \mathbb{E}[n_k(t)] \) as

\[
\mathbb{E}[n_k(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{k_t = k\}}\right] 
\]

\[
= \sum_{t=0}^{T-1} \Pr(k_{t+1} = k) 
\]

\[
= \sum_{t=1}^{K_{t_0}} \Pr(k_t = k) + \sum_{t=K_{t_0}}^{T-1} \Pr(k_{t+1} = k) 
\]

\[
\leq K_{t_0} + \sum_{t=K_{t_0}}^{T-1} \Pr(k_{t+1} = k, n_{k^*}(t) = \max_{k'} n_{k'}(t)) + \sum_{t=K_{t_0}}^{T-1} \sum_{t'=k_{t_0}^* \neq k^*} \left( \Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \times \Pr(k_{t+1} = k|n_{k'}(t) = \max_{k''} n_{k''}(t)) \right) 
\]

\[
\leq K_{t_0} + \sum_{t=K_{t_0}}^{T-1} \Pr(k_{t+1} = k, n_{k^*}(t) = \max_{k'} n_{k'}(t)) + \sum_{t=K_{t_0}}^{T-1} \sum_{t'=k_{t_0}^* \neq k^*} \Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) 
\]

\[
\leq K_{t_0} + \sum_{t=K_{t_0}}^{T-1} 2t^{1-\alpha} + \sum_{t=K_{t_0}}^{T} \sum_{t'=k_{t_0}^* \neq k^*} \Pr\left(n_{k'}(t) \geq \frac{t}{K}\right) 
\]

\[
\leq K_{t_0} + \sum_{t=1}^{T} 2K t^{1-\alpha} + K^2(K-1) \sum_{t=K_{t_0}}^{T} 6 \left(\frac{t}{K}\right)^{2-\alpha}. 
\]

**Proof of Theorem 1** For any suboptimal arm \( k \neq k^* \),

\[
\mathbb{E}[n_k(T)] \leq \sum_{t=1}^{T} \Pr(k_t = k) 
\]

\[
= \sum_{t=1}^{T} \Pr((k_t = k, E_1(t)) \cup (E^c_1(t), k_t = k)) 
\]

\[
\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \sum_{t=1}^{T} \Pr(E^c_1(t), k_t = k) 
\]

\[
\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \sum_{t=0}^{T-1} \Pr(I_k(t) > I_{k^*}(t), k_{t+1} = k) 
\]

\[
= \sum_{t=1}^{T} 2K t^{1-\alpha} + \sum_{t=0}^{T-1} \Pr(I_k(t) > I_{k^*}(t), k_{t+1} = k) 
\]

\[
\leq 8\alpha^2 \frac{2 \log(T)}{\Delta_k^2} + \frac{2\alpha}{\alpha - 2} + \sum_{t=1}^{T} 2K t^{1-\alpha}. 
\]

Here, (65) follows from Lemma 3. We have (66) from the analysis of UCB for the classical bandit problem for details see proof of Theorem 2.1 in [24].

**Proof of Theorem 3** Follows directly by combining the results on Theorem 1 and Theorem 2.

### C Reward functions for the experiments

As mentioned in Section 6, we use a part of the Movielens dataset (50%) as the training dataset, on which the mean reward mappings from meta-users to different genres are learned. Figure 11 represents the learned mappings on the training dataset.

Here, (68) follows from Lemma 5 and (59) follows from Lemma 6.
Figure 11: Learned reward mappings from 106 meta-users to each of the movie genres, i.e., the $\mu_k(\theta)$ in the problem setup, with $\theta$ representing different meta-users and $k(\text{arm})$ representing different movie genres.