Exploiting Correlation in Finite-Armed Structured Bandits

Samarth Gupta
Carnegie Mellon University
Pittsburgh, PA 15213

Gauri Joshi
Carnegie Mellon University
Pittsburgh, PA 15213

Osman Yağan
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We consider a correlated multi-armed bandit problem in which rewards of arms are correlated through a hidden parameter. Our approach exploits the correlation among arms to identify some arms as sub-optimal and pulls 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; explicit conditions needed for bounded regret to be possible are also provided by analyzing regret lower bounds. We propose several variants of our approach that generalize classical bandit algorithms such as UCB, Thompson sampling, KL-UCB to the structured bandit setting, and empirically demonstrate their superiority via simulations.

1 Introduction

The Multi-armed bandit problem [1] (MAB) falls under the umbrella of sequential decision-making problems. 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 $k \in \mathcal{K}$ and receives a random reward $R_{k,t}$ with unknown mean μ_k. The goal of the player is to maximize her cumulative reward. In order to do so, the player must balance exploration and exploitation of arms. This classical K-armed bandit formulation assumes independence of the rewards of different arms. However in many online learning problems, such as dynamic pricing and drug dosage optimization, there is a correlation between rewards of different actions.

Motivated by this, we consider a correlated multi-armed bandit problem, in which the mean reward of different arms are related through a common hidden parameter θ. Specifically, the expected return of arm k, i.e., $\mathbb{E}\left[R_{k,t}|k,\theta\right] = \mu_k(\theta)$. In the setting considered, the mean reward functions, $\mu_k(\theta)$, are known to the player but the true value of shared parameter θ^* is unknown. The dependence on the common parameter introduces a structure in this MAB problem. This makes the model interesting as the rewards observed from an arm can provide information about mean rewards from other arms. Similar models have been considered in [2, 3, 4], but as explained in Section 2.2, we consider a more general setting that subsumes the models in [2, 3, 5].

There are many 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 $\mathcal{P}$, and the average revenue in time slot t is a function of the selected price p_t and the market size θ. These functions are typically known from literature [6], 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. Authors in [4] provide a similar example for the purpose of advertising, in which a company needs to decide

Preliminary work. Under review by AISTATS 2019. Do not distribute.

what form of advertising to purchase so as to maximize its profit. The problem set-up is also relevant in system diagnosis, where θ represents the unknown cause of a failure in the system, and μ_k's represent the response of system to different actions. Other applications of this model include cellular coverage optimization [7] and drug dosage optimization [3]. Our general treatment of the structured bandit setting will allow our work to be helpful in all these problems.

Main Contributions.

1) We consider a general setting for the problem which subsumes previously considered models. [2, 3, 5].

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 θ 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) Our finite-time regret analysis reveals how this idea leads to a smaller regret as compared to 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 $K-1$ arms are pulled $\mathrm{O}(\log T)$ times. The value of C can be much less than K and can even be 1, in which case our proposed algorithm achieves bounded regret! Our analysis reveals that proposed 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 [8], KL-UCB [9] etc.) in the structured bandit setting. This extension was deemed to be not possible easily for the UCB-S algorithm proposed in [4].

5) We design two variants of UCB-C, namely UCB-int and UCB-min, and demonstrate the empirical superiority of the proposed algorithms in different scenarios.

2 Problem Formulation

2.1 System Model

We consider a K-armed bandit setting in which the rewards of arms $\{1, 2 \ldots, K\}$ are correlated. As shown in Figure 1, we assume that the mean reward of each arm is dependent on a common hidden parameter θ. At each time step t, the player pulls arm $k_t \in \{1, 2 \ldots K\}$ and observes the reward $R_{k_t,t}$.

The reward $R_{k_t,t}$ obtained at time step t is a random variable with mean $\mu_{k_t}(\theta)$, where θ is a fixed unknown parameter which lies in a known set Θ; our formulation

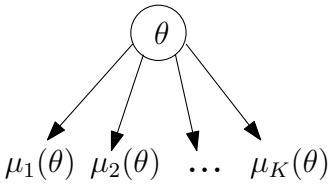

Figure 1: *Structured bandit setup: rewards of different arms are correlated through hidden parameter θ.*

allows the set Θ to be a countable or uncountable. The functions $\{\mu_1, \mu_2, \ldots, \mu_K\} : \Theta \to \mathbb{R}$ are known to the player but the true value of parameter θ, i.e., $\theta^* \in \Theta$, is unknown. The parameter θ can also be a vector. The objective of the player is to maximize her cumulative reward in T rounds. If a player had known the true value θ^*, then she would always pull the arm having the highest mean reward for the parameter θ^*, as that would lead to maximum cumulative reward in expectation. Motivated by this, we call the optimal arm as $k^* = \arg\max_{k \in \mathcal{K}} \mu_k(\theta^*)$, i.e., the best arm for the true parameter θ^*. The sub-optimality gap of arm k, Δ_k, is defined as the difference between mean reward of the optimal arm and of arm k; i.e., $\Delta_k \triangleq \mu_{k^*}(\theta^*) - \mu_k(\theta^*)$. The performance of a player is evaluated by the cumulative regret defined as:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \mu_{k^*}(\theta^*) - \mu_{k_t}(\theta^*) = \sum_{t=1}^{T} \Delta_{k_t} = \sum_{k \neq k^*} n_k(T) \Delta_k.$$

Here $n_k(T)$ is a random variable denoting the number of times arm k is pulled in a total of T time slots. The cumulative regret quantifies the performance of a player in comparison to an oracle that pulls the optimal arm at each time slot. Thus, the smaller is the regret, the better is the performance of the player.

As mentioned earlier, the player only knows the *mean* reward functions $\mu_k(\theta) = \mathbb{E}\left[R_{k,t} | \theta, k\right]$ and not the conditional distribution of rewards i.e., $p(R_{k,t}|\theta, k)$ is not known. Throughout the paper, we assume that the rewards $R_{k,t}$ are sub-Gaussian with variance proxy σ^2, i.e., $\mathbb{E}\left[\exp\left(s\left(R_{k,t} - \mathbb{E}\left[R_{k,t}\right]\right)\right)\right] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right)\ \forall s \in \mathbb{R}$, and σ is known to the player. Both of these assumptions are common in the multi-armed bandit literature [4, 10, 11, 12, 13]. In particular, the sub-Gaussianity of rewards enables us to apply Hoeffding's inequality, which is essential for the regret analysis.

We would like to highlight that we make no assumptions on the functions $\{\mu_1, \mu_2, \ldots, \mu_K\}$, unlike some previous works ([2, 3, 5]) that place restrictive assumptions on the functions. Due to the general nature of our setup, our model subsumes these previously studied frameworks and is applicable to much more general sce-

narios as well. The similarities and differences between our model and existing studies are discussed next.

2.2 Connections with Previously Studied Bandit Models

Classical MAB. Under the classical Multi-armed bandit setting, the rewards obtained from each arm are independent. By considering $\theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$ and $\mu_k = \theta_k$, our setting reduces to the classical MAB setting. Our proposed algorithm will in fact perform UCB/Thompson sampling ([1, 8]) in this special case.

Global Bandits [2]. In [2], a model where mean reward functions are dependent on a common scalar parameter in studied. A key assumption in [2] 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 proposed algorithm also achieves bounded regret. Hence, our formulation covers the setting described in [2].

Regional Bandits [3]. The paper [3] studies a setting in which there are M common unknown parameters. The mean reward function of each arm depends on one of these M parameters, θ_m. These mean reward functions of each arm are assumed to be invertible and Hölder-continuous as a function of θ_m. The setting described in [3] is captured in our formulation by setting $\theta = (\theta_1, \theta_2, \ldots, \theta_M)$. In fact, our problem setup allows for the mean reward function of arm k to be a function of a combination of all of these M parameters and these mean reward functions need not be invertible.

Structured bandits with linear functions [5]. In [5], the authors consider a model in which rewards of all arms depend on a common parameter. However, they assume that the mean reward functions, $\mu_k(\theta)$ are linear functions of θ. Under this assumption, they design a greedy policy that achieves bounded regret. Our formulation places no such restriction on the reward functions, and thus is more general. In the specific cases where reward functions are linear, our proposed algorithm also achieves bounded regret.

Finite-armed generalized linear bandits [14]. Under the finite-armed linear bandit setting [14], the reward function of arm x_k is $\theta^\intercal x_k$. Here, θ is the shared unknown parameter. Similarly, when $\mu_k(\theta) = g(\theta^\intercal x_k)$, this becomes the generalized linear bandit setting [15], for some known function g. One can easily see that, our setting perfectly captures both of them.

Minimal Exploration in Structured Bandits [16]. Authors in [16] consider a problem formulation

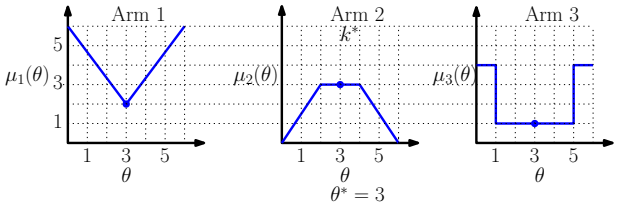

Figure 2: *An example showing a 3-armed bandit problem. The figures show the mean rewards of the three arms as a function of θ. Since the true parameter $\theta* = 3$, Arm 2 is the optimal arm while Arms 1 and 3 are sub-optimal.*

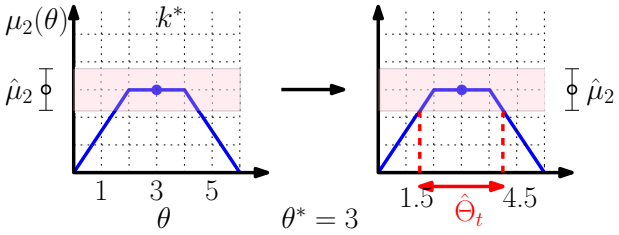

Figure 3: *The samples of Arm 2, can be used to obtain empirical mean of arm 2. The empirical mean can be used to construct a region (shaded pink) within which $\mu_2(\theta^*)$ lies with high probability. This region can be used to identify a the high confidence set Θ_t that contains the true parameter θ^* with high probability.*

that is more general than the setting described in this paper. However, the focus of [16] is to obtain asymptotically optimal regret for the regimes when regret scales as $\log(T)$. When all arms are *non-competitive*, the solution to the optimization problem described in [16, Theorem 1] becomes 0, causing the algorithm to get stuck in the exploitation phase and not perform properly in such settings. Moreover, they assume $\theta \to \mu_k(\theta)$ is continuous, while we make no such assumption.

Finite-armed structured bandits [4]. The work closest to ours is [4]. The authors in [4] consider the same model that we consider and propose the UCB-S algorithm, which is a UCB-style algorithm for this setting. We take a different approach to this problem, and propose a novel algorithm that separates *implicit* and *explicit* exploration, that allows us to extend our UCB style algorithm to other classical bandit algorithms such as Thompson sampling. A Thompson sampling style algorithm was not proposed in [4]. Through simulations, we make comparisons of our proposed algorithms against the UCB-S algorithm proposed in [4].

2.3 Intuitions for developing an algorithm

Classic multi-armed bandit algorithms such as UCB [1] and Thompson sampling [8] rely on *explicit* exploration

of *empirically sub-optimal* arms to learn the optimal action. In our framework, since mean rewards of all arms are dependent on a common parameter, obtaining an estimate of θ^* from the samples observed till slot t can give us some information on the mean rewards of all arms. This additional knowledge can then be used to reduce the *exploration* needed when designing bandit algorithms. Identifying sub-optimal arms through this estimate of θ^* can be thought of as *implicit* exploration.

Consider the example shown in Figure 2. In this case, the true parameter, θ^*, is equal to 3, and the mean reward of Arm 1 is 2, Arm 2 is 3, and that of Arm 3 is 1. Thus, the optimal arm in this setup is Arm 2. Assume now that the player has obtained a large number of samples of arm 2 at a given time step. Based on the samples observed from Arm 2, the player has an empirical estimate of the mean reward as

$$\hat{\mu}_2(t) = \frac{\sum_{\tau=1}^{t} R_{k_\tau,\tau} \mathbb{1}_{k_\tau=2}}{n_2(t)}. \tag{1}$$

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 $\hat{\Theta}_t$ within which the true parameter θ^* lies with high probability. For example, in Figure 3 that region is the set $[1.5, 4.5]$. Upon identifying this set, we can now see that if θ^* indeed lies in this set, then Arm 3 cannot be optimal as it is sub-optimal compared to Arm 2 for all values of $\theta \in [1.5, 4.5]$. 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 θ in this set.

We formalize this idea of identifying *non-competitive* arms into an online algorithm that performs both implicit and explicit exploration. The proposed algorithm, presented in the next section, successfully reduces a K armed bandit problem to a C armed bandit problem, where $C \leq K$ is the number of *competitive* arms, defined formally in Section 3. More interestingly, this algorithm can lead to *bounded* (i.e., not scaling with T) regret in certain regimes 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

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

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 $k_{t+1} = \arg\max_{k \in \mathcal{K}} S_{k,t}$ at time step t. Here, $S_{k,t}$ is the sample obtained from the posterior distribution of μ_k, i.e.,

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

Since mean rewards are correlated through the hidden parameter θ^* in the structured bandit model, obtaining an estimate of θ^* can help identify the optimal arm. In our approach, we will identify subset of arms, called the *competitive* arms, through the estimate of θ^* and then perform UCB or TS over that set of arms. We now define the notion of *competitive* and *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 $\hat{\Theta}_t$. The set $\hat{\Theta}_t$ represents the set of values in which the true parameter θ^* lies with high confidence, based on rewards observed until time t. Next, we define the notions of $\hat{\Theta}_t$-Competitive and $\hat{\Theta}_t$-Non-competitive arms.

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

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

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

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

We now introduce the notion of ϵ-non-competitive arm.

Definition 3 (ϵ-non-competitive arm)**.** *We call an arm k as ϵ-non-competitive if*

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

Informally, this means that if an arm is ϵ-non-competitive, then it is $\hat{\Theta}$-Non-competitive, with $\hat{\Theta}$ being the set of θ that do not change the true mean of the optimal arm by more than ϵ.

Throughout, we say that an arm k is *competitive* if there is no $\epsilon > 0$ for which it is ϵ-non-competitive. The set of all competitive arms is denoted by $\mathcal{C}$ and the number of competitive arms by $C = |\mathcal{C}|$.

Let Θ^* be defined as the set $\Theta^* = \{\theta : \mu_{k^*}(\theta) = \mu_{k^*}(\theta^*)\}$. We can view Θ^* as the confidence set Θ_t after the optimal arm is sampled infinitely many times. This definition ensures that if an arm is Θ^*-Competitive, then it is competitive.

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 (say 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.

Constructing a confidence set, $\hat{\Theta}_t$. From the samples observed till time step t, we identify the arm that has been selected the maximum number of times so far, namely $k^{\max} = \arg\max_{k\in\mathcal{K}} n_k(t)$. We define the confidence set as follows:

$$\hat{\Theta}_t = \left\{ \theta :| \mu_{k^{\max}}(\theta) - \hat{\mu}_{k^{\max}}(t) |< \sqrt{\frac{2\alpha\sigma^2 \log t}{n_{k^{\max}}(t)}} \right\}.$$

Here $\hat{\mu}_{k^{\max}}$ is the empirical mean of rewards obtained in $n_{k^{\max}}$ samples of arm $k^{\max}$. We construct the arm with samples of $k^{\max}$ as it has smallest variance in its empirical estimate among all arms. In our regret analysis, we show that using samples of just one arm suffices to achieve the desired dimension reduction and bounded regret properties. In Section 5, we present and discuss the UCB-int algorithm that constructs the confidence set using samples of all arms.

Identifying $\hat{\Theta}_t$-Non-competitive arms. At each time step $t+1$, we define the set $\mathcal{C}_t$ as the set of $\hat{\Theta}_t$-Competitive arms, that includes all arms k that satisfy $\mu_k = \max_{\ell\in\mathcal{K}} \mu_\ell(\theta)$ for some $\theta \in \hat{\Theta}_t$. The rest of the arms, termed as $\hat{\Theta}_t$-Non-competitive, are eliminated for round $t+1$ and are not considered in the next part of the algorithm.

Algorithm 1 UCB-C Correlated UCB Algorithm

Input: Reward Functions $\{\mu_1, \mu_2 \dots \mu_K\}$
Initialize: $n_k = 0, I_k = \infty$ for all $k \in \{1, 2, \dots K\}$

for each round $t+1$ **do**
  Find $k^{\max} = \arg\max_k n_k(t)$, the arm that has been pulled most times until round t
  Confidence set construction:

$$\hat{\Theta}_t = \left\{ \theta :| \mu_{k^{\max}}(\theta) - \hat{\mu}_{k^{\max}}(t) |< \sqrt{\frac{2\alpha\sigma^2 \log t}{n_{k^{\max}}(t)}} \right\}.$$

  Define competitive set $\mathcal{C}_t$:

$$\mathcal{C}_t = \left\{ k : \mu_k(\theta) = \max_{\ell\in\mathcal{K}} \mu_\ell(\theta) \quad \text{for some } \theta \in \Theta_t \right\}.$$

  UCB among competitive arms

$$k_{t+1} = \arg\max_{k\in\mathcal{C}_t} \hat{\mu}_k(t) + \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}.$$

  Update empirical mean,$\hat{\mu}_k$, UCB indices, I_k, and n_k for every arm k.
end for

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 $t+1$. For example, in the case of UCB-C, the next arm k_{t+1} is selected as $\arg\max_{k\in\mathcal{C}_t} I_k(t)$, with $I_k(t) = \hat{\mu}_k(t) + \sqrt{\frac{2\alpha\sigma^2}{n_k(t)}}$.

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 [17, 18, 19, 20, 21] 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 [4].

The idea of eliminating *non-competitive* arms was initially proposed in [22] for studying multi-armed bandits with a latent random source. However, given the different nature of the problem studied in [22], an entirely different definition of arm *competitiveness* was used.

Algorithm 2 TS-C Correlated Thompson sampling

Input: Reward Functions $\{\mu_1, \mu_2 \dots \mu_K\}$
Initialize: $n_k = 0$ for all $k \in \{1, 2, \dots K\}$
for each round $t+1$ **do**
  Find $k^{\max} = \arg\max_k n_k(t)$, the arm that has been pulled most times until round t
  Confidence set construction:

$$\hat{\Theta}_t = \left\{ \theta :\mid \mu_{k^{\max}}(\theta) - \hat{\mu}_{k^{\max}}(t) \mid < \sqrt{\frac{2\alpha\sigma^2 \log t}{n_{k^{\max}}(t)}} \right\}.$$

  Define competitive set $\mathcal{C}_t$:

$$\mathcal{C}_t = \left\{ k : \mu_k(\theta) = \max_{\ell \in \mathcal{K}} \mu_\ell(\theta) \quad \text{for some} \theta \in \Theta_t \right\}.$$

  Apply Thompson sampling on $\mathcal{C}_t$
  for $k \in \mathcal{C}_t$ **do**
    Sample $S_{k,t} \sim \mathcal{N}\left(\hat{\mu}_k(t), \frac{\sigma^2}{n_k(t)}\right)$.
  end for
  $k_{t+1} = \arg\max_{k \in \mathcal{C}_t} S_{k,t}$
  Update empirical mean,$\hat{\mu}_k$ and n_k for all arm k.
end for

4 Regret Analysis and Bounds

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

$$\mathbb{E}\left[\mathrm{Reg}(T)\right] = \sum_{k=1}^{K} \mathbb{E}\left[n_k(T)\right] \Delta_k. \tag{2}$$

Here, $\Delta_k = \mu_{k^*}(\theta^*) - \mu_k(\theta^*)$ and $n_k(T)$ is the number of times arm k is pulled in a total of T time steps.

To analyze the expected regret of a proposed algorithm (as given by (2) above), we need to determine $\mathbb{E}\left[n_k(T)\right]$ for each sub-optimal arm $k \neq k^*$. Our first result shows that expected pulls for any arm is $\mathrm{O}(\log T)$.

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

$$\mathbb{E}\left[n_k(T)\right] \leq 8\alpha\sigma^2 \frac{\log T}{\Delta_k^2} + \frac{2\alpha}{\alpha - 2} + \sum_{t=1}^{T} 2t^{1-\alpha} \tag{3}$$

Our next result shows that the expected number of pulls for an ϵ_k-non-competitive arm are bounded.

Theorem 2 (Expected pulls of Non-competitive Arms)**.** *If an arm k is ϵ_k-non-competitive, then the number of times it is pulled by UCB-C is upper bounded as*

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

where,

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

Plugging the results of Theorem 1 and Theorem 2 in (2) yields the following bound on the expected regret of UCB-C Algortihm.

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

$$\begin{aligned}\mathbb{E}\left[Reg(T)\right] &\leq \sum_{k \in \mathcal{C}\setminus\{k^*\}} \Delta_k U_k^{(c)}(T) + \sum_{\ell \in \mathcal{K}\setminus\mathcal{C}} \Delta_\ell U_\ell^{(nc)}(T), \\ &= (C-1)\cdot \mathrm{O}(\log T) + \mathrm{O}(1),\end{aligned}$$

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

Dimension Reduction. The classic UCB algorithm that is agnostic to the structure of the problem pulls each of the $(K-1)$ sub-optimal arms $\mathrm{O}(\log T)$ times. In contrast, our algorithm pulls only $(C-1)$ sub-optimal arms $\mathrm{O}(\log T)$ times, where $C \leq K$. In fact, when $C = 1$, all sub-optimal arms are pulled only $\mathrm{O}(1)$ times, leading to a bounded regret. Such cases can arise quite often in practical settings. For example, when the optimal arm k^* is invertible around θ^*, the set Θ^* 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 [16].

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

$$\liminf_{T\to\infty} \frac{Reg(T)}{\log T} \geq L(\theta), \ \textit{where}$$

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

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

The proof of this proposition, given in Appendix, follows from a lower bound derived in [16]. This lower bound leads us to the following observation.

Remark 1 (Bounded regret whenever possible)**.** *The result on lower bound in Proposition 1 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 have dimensionality reduction for cases when regret is logarithmic.*

5 Variants of UCB-C

Recall that the design of UCB-C and TS-C algorithms involved the construction of the confidence set $\hat{\Theta}_t$. This confidence set was constructed by selecting all θ for which $\mu_{k^{\max}}(\theta)$ is inside a ball of size $\sqrt{\frac{2\alpha\sigma^2 \log T}{n_{k^{\max}}(T)}}$ around the empirical mean $\hat{\mu}_{k^{\max}}$. In this section, we discuss two other methods of constructing the confidence set to show that our idea of separating *implicit* and *explicit* exploration can be easily extended to design new algorithms. These extensions can lead to lower empirical regret than the UCB-C algorithm in some cases. We now describe the two algorithms and evaluate their regret bounds.

5.1 The UCB-int Algorithm

The Algorithm UCB-C, constructs the confidence set $\hat{\Theta}_t$ using just the samples of arm $k^{\max}$. We now present the UCB-int Algorithm, which differs from UCB-C in this aspect. At an additional computation cost, it constructs the confidence set $\hat{\Theta}_t$ by using samples of all the arms pulled so far. More specifically, UCB-int constructs $\hat{\Theta}_t$ as follows:

$$\hat{\Theta}_t = \left\{ \theta : \forall k \in \mathcal{K} \quad |\mu_k(\theta) - \hat{\mu}_k| < \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}} \right\}. \tag{4}$$

Similar to UCB-C, the UCB-int Algorithm works for any functions $\mu_k(\theta)$ and any set Θ. The UCB-int also has performance guarantees similar to UCB-C, i.e., $Reg(T) = (C-1)\mathrm{O}(\log T) + \mathrm{O}(1)$. It also enjoys the same property of achieving bounded regret whenever possible. We present the exact regret bound of UCB-int in the Appendix. Since we consider samples of all arms in constructing the confidence set for UCB-int, the confidence set Θ_t is smaller for UCB-int than for the UCB-C. Hence, UCB-int is more aggressive in removal of non-competitive arms and can obtain better empirical performance than the UCB-C algorithm.

5.2 The UCB-min Algorithm

The Algorithm UCB-C, was designed to accommodate all type of functions $\mu_k(\theta)$ and all sets Θ. In this Section we design an aggressive algorithm, named UCB-min, for a class of structured bandit settings. The problem settings for which UCB-min is designed includes all cases where Θ is a countable set.

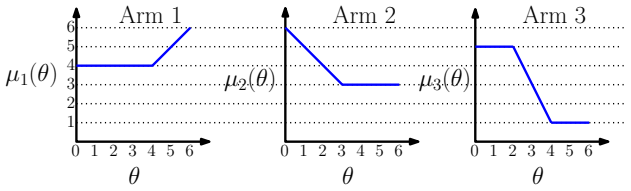

Figure 4: *An example where Arm 2 is optimal for $\theta^* \in [0, 1]$, Arm 3 is optimal for $\theta^* \in [1, 2.5]$ and Arm 1 is optimal for $\theta^* \in [2.5, 6]$.*

The UCB-min Algorithm has the same three components as that of UCB-C. It differs from UCB-C only in the first component, i.e., the construction of the confidence set. Instead of considering a ball of size $\sqrt{\frac{2\alpha\sigma^2 \log T}{n_{k^{\max}}(T)}}$ around $\hat{\mu}_{k^{\max}}$, we directly choose $\hat{\Theta}_t$ by selecting θ corresponding to the closest $\mu_{k^{\max}}(\theta)$ value. More specifically, we construct $\hat{\Theta}_t$ as follows:

$$\mu^{\min} = \min_{\theta \in \Theta} |\mu_{k^{\max}}(\theta) - \hat{\mu}_{k^{\max}}|. \tag{5}$$

$$\hat{\Theta}_t = \{\theta : \mu_{k^{\max}}(\theta) = \mu^{\min}\}. \tag{6}$$

The UCB-min algorithm is designed for structured bandit settings that satisfy the following assumption.

Assumption 1. *There exists $\delta > 0$ such that for all sub-optimal arms $k \neq k^*$,*

$$\mu_{k^*}(\theta) = \max_{\ell \in \mathcal{K}} \mu_\ell(\theta) \text{ for all } \theta : |\mu_k(\theta^*) - \mu_k(\theta)| < \delta.$$

Informally this means that the optimal arm at θ^* remains $\hat{\Theta}_t$-Competitive as long as $\mu^{\min}$ lies at a distance of δ away from $\mu_{k^{\max}}(\theta^*)$. Assumption 1 is always satisfied when Θ is a countable set. Under this assumption, UCB-min enjoys similar regret guarantees as that of UCB-C. We have $Reg(T) = (C-1)\mathrm{O}(\log T) + \mathrm{O}(1)$, with C denoting the number of competitive arms. As was the case with UCB-C, it also achieves bounded regret whenever possible. The regret upper bound for UCB-min, which depends on δ, is given in the appendix.

6 Simulation Results

We now study the empirical performance of the proposed algorithms. For all simulations we choose $\alpha = 3.5$. Rewards are drawn from the distribution $\mathcal{N}(\mu_k(\theta^*), 4)$, i.e., $\sigma = 2$. In each result we average the regret over 100 experiments. We first show how UCB-C is able to achieve dimension reduction.

Dimension Reduction. In Figure 4 we compare the regret of the UCB-C algorithm with classic UCB for

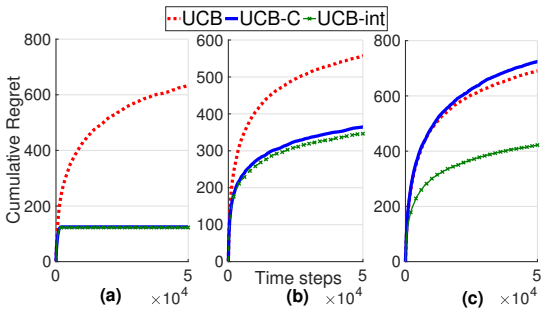

Figure 5: *Cumulative regret of UCB, UCB-C and UCB-int for the setting in Figure 4. The true parameter $\theta^* = 0.5$ in (a), $\theta^* = 1.8$ in (b) and $\theta^* = 2.8$ in (c).*

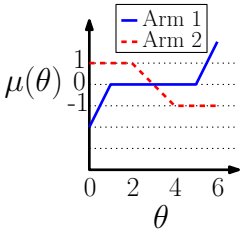

Figure 6: *Arm 2 is optimal for $\theta^* \in [0, 3]$ and Arm 1 is optimal for $\theta^* \in [3, 5]$.*

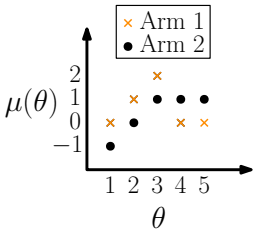

Figure 7: *Arm 1 is optimal for $\theta^* \in \{1, 2, 3\}$ and Arm 2 is optimal otherwise.*

the example considered in Figure 4. 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 Algorithms UCB-C and UCB-int achieve 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 and UCB-int to pull Arm 1 only O(1) times due to which we notice significantly reduced regret with UCB-C and UCB-int 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. Since UCB-C performs UCB over the set of $\hat{\Theta}_t$-Competitive arms at each round, its performance is similar to that of the UCB algorithm in this case. The UCB-int algorithm uses samples from all arms to generate the confidence set Θ_t, which helps in achieving empirically smaller regret for this setting.

Comparison with UCB-S. We now compare the performance of our UCB-C and TS-C algorithms against the UCB and UCB-S Algorithm proposed in [4]. We consider the example shown in Figure 6. For the situations in which the parameter $\theta^* < 3$, Arm 2 is optimal and Arm 1 is non-competitive. When θ^* takes values in $[3, 5]$, Arm 1 is the optimal arm while Arm 2 is competitive. We plot the cumulative regret of the UCB, UCB-S, UCB-C and TS-C algorithms over 50000 time steps for the values of θ^* between 0 and 5 in Figure 8. When

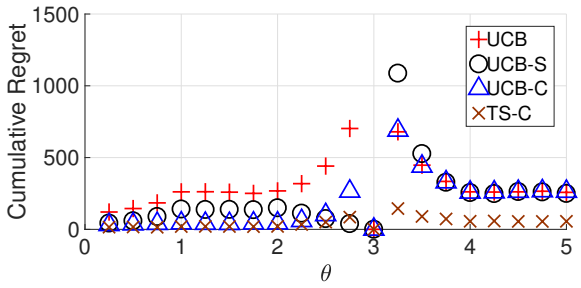

Figure 8: *Cumulative regret of UCB,UCB-S,UCB-C and TS-C for the Example in Figure 6 over 50000 runs compared over different values of θ^**

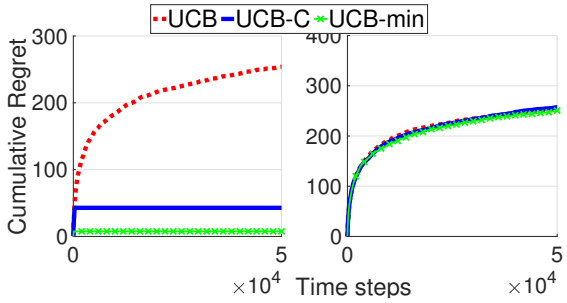

Figure 9: *Cumulative regret of UCB, UCB-C and UCB-min for the example considered in Figure 7 with $\theta^* = 2$ in (a) and $\theta^* = 4$ in (b).*

θ^* is below 3, UCB-S, UCB-C and TS-C all obtain lesser regret than UCB as they are able to identify the sub-optimal arm as non-competitive. When $\theta^* \in (3, 5)$, we see that 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 the algorithm in UCB-S selects $k_{t+1} = \arg\max_{k \in \mathcal{K}} \sup_{\theta \in \Theta_t} \mu_k(\theta)$. The sup causes the algorithm to prefer certain arms over others, this is the reason for its large regret at $\theta^* = 3.25$ and relatively smaller regret at $\theta^* = 2.75$. We notice that TS-C achieves significantly less regret as compared to other algorithms as Thompson sampling can offer significant empirical improvement over UCB. This highlights that the possibility to incorporate Thompson sampling in algorithm is beneficial, this extension to Thompson sampling was not possible in [4].

Performance of UCB-min. We now compare the performance of the UCB-min algorithm against the UCB-C and UCB algorithm for the example considered in Figure 7. In this example, Θ is a countable set, which allows us to use the UCB-min algorithm here. The aggressive nature of UCB-min in selection of Θ_t helps it to achieve smaller bounded regret than UCB-C, as shown in Figure 9 where $\theta^* = 2$.

7 Concluding Remarks

In this work, we studied a correlated bandit problem in which the rewards of different arms are correlated through a common shared parameter. By using reward samples of an arm, we were able to generate estimates on mean reward of other arms. This approach allowed us to identify some sub-optimal arms without having to explore them explicitly. The finite time regret analysis of the proposed UCB-C algorithm reveals that it is able to reduce the K-armed bandit problem to a C-armed bandit problem. In addition, we showed that UCB-C achieves bounded regret whenever possible. Ongoing work includes the finite-time regret analysis of TS-C algorithm. An interesting future direction is to study this problem when the number of arms is *large*. We also plan to study the best-arm identification version of this problem.

References

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

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134, 2013.

[12] Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pages 638–646, 2013.

[13] Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *International Conference on Machine Learning*, pages 136–144, 2014.

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

[15] 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.

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

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

[18] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

[19] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

[20] 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.

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

[22] Samarth Gupta, Gauri Joshi, and Osman Yağan. Correlated multi-armed bandits with a latent random source. *arXiv preprint arXiv:1808.05904*, 2018.

[23] 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 [16] to state the Proposition 1.

Theorem 4 (Lower bound, Theorem 1 in [16].)**.** *For any uniformly good algorithm [1], and for any $\theta \in \Theta$, we have:*

$$\liminf_{T\to\infty} \frac{Reg(T)}{\log T} \geq L(\theta),$$

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

$$\min_{\eta(k)\geq 0, k\in\mathcal{K}} \sum_{k\in\mathcal{K}} \eta(k) \left(\max_{\ell\in\mathcal{K}} \mu_\ell(\theta) - \mu_k(\theta)\right)$$
$$\text{subject to} \sum_{k\in\mathcal{K}} \eta(k) D(\theta,\lambda,k) \geq 1, \forall \lambda \in \Lambda(\Theta), \tag{7}$$

$$\text{where} \quad \Lambda(\theta) = \{\lambda \in \Theta^* : k^* \neq \arg\max_{k\in\mathcal{K}} \mu_k(\lambda)\}. \tag{8}$$

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, π, is uniformly good if $Reg^\pi(T,\theta) = o(T^a)$ for all $a > 0$ and all $\theta \in \Theta$.

We see that the solution to the optimization problem (7) 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.

B Results for UCB-int and UCB-min

B.1 Regret bounds for UCB-int

We state the regret bounds for UCB-int Algorithm here.

Theorem 5 (Expected pulls for any arm)**.** *Expected number of times any arm is pulled by UCB-int Algorithm is upper bounded as*

$$\mathbb{E}\left[n_k(T)\right] \leq 8\alpha\sigma^2 \frac{\log T}{\Delta_k^2} + \frac{2\alpha}{\alpha-2} + \sum_{t=1}^{T} 2Kt^{1-\alpha} \tag{9}$$

Theorem 6 (Expected pulls for non-competitive arms)**.** *The Expected number of times an ϵ_k non-competitive arm is pulled by UCB-int Algorithm is upper bounded as*

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

with,

$$t_0 = \inf\Bigg\{\tau \geq 2 : \Delta_k \geq 4\sqrt{\frac{K\alpha\sigma^2\log\tau}{\tau}}, \\ \epsilon_k \geq \sqrt{\frac{8\alpha\sigma^2 K\log\tau}{\tau}}\Bigg\} \tag{10}$$

Combining the above two results gives us the following regret bound for the UCB-int Algorithm.

Theorem 7 (Regret upper bound)**.** *For $\alpha > 3$, the expected regret of the UCB-int Algorithm is upper bounded as,*

$$\mathbb{E}\left[Reg(T)\right] \leq \sum_{k\in\mathcal{C}\setminus\{k^*\}} \Delta_k U_k^{(c)}(T) + \sum_{\ell\in\mathcal{K}\setminus\mathcal{C}} \Delta_\ell U_\ell^{(nc)}(T), \tag{11}$$

$$= (C-1)\cdot \mathrm{O}(\log T) + \mathrm{O}(1), \tag{12}$$

Here, $U_k^{(c)}(T)$ is the upper bound on $\mathbb{E}\left[n_k(T)\right]$ given in Theorem 5 and $U_\ell^{(nc)}(T)$ is the upper bound on $\mathbb{E}\left[n_\ell(T)\right]$ given in Theorem 6. The set $\mathcal{C}$ denotes the set of competitive arms and C is the cardinality of that set.

Observe that this regret bound is similar to the regret bound for UCB-C algorithm and enjoys the same properties of achieving bounded regret whenever possible and performing dimension reduction.

B.2 Regret bounds for UCB-min

We now present the performance guarantees of the UCB-min Algorithm. Under the Assumption 1 and t_0 defined as

$$t_0 = \inf\Bigg\{\tau \geq 2 : \Delta_{\min} \geq 4\sqrt{\frac{K\alpha\sigma^2\log\tau}{\tau}}, \tag{13}$$

$$\delta \geq \sqrt{\frac{8K\alpha\sigma^2\log\tau}{\tau}}\Bigg\} \tag{14}$$

We have the following result on expected pulls of arms,

Theorem 8 (Expected pulls for any arm)**.** *Expected number of times any arm is pulled is upper bounded in the UCB-min Algorithm as*

$$\mathbb{E}\left[n_k(T)\right] \leq 8\alpha\sigma^2\frac{\log(T)}{\Delta_k^2} + \frac{2\alpha}{\alpha-2} + \sum_{t=1}^{T} 2t\exp\left(-\frac{\delta^2 t}{8K\sigma^2}\right). \tag{15}$$

Theorem 9 (Expected pulls for ϵ_k*-non-competitive* arms)**.** *Expected number of times an ϵ_k-non-competitive arm is pulled by the UCB-min Algorithm is upper bounded as*

$$\mathbb{E}\left[n_k(t)\right] \leq Kt_0 + \sum_{t=1}^{T} 2t\exp\left(-\frac{\epsilon_k^2 t}{8K\sigma^2}\right) + K(K-1)\sum_{t=Kt_0}^{T} 6\left(\frac{t}{K}\right)^{2-\alpha} \tag{16}$$

Combining the above two result yields us the following regret bound for UCB-min Algorithm.

Theorem 10 (Regret upper bound)**.** *For $\alpha > 3$, the expected regret of the UCB-min Algorithm is upper bounded as,*

$$\mathbb{E}\left[Reg(T)\right] \leq \sum_{k\in\mathcal{C}\setminus\{k^*\}} \Delta_k U_k^{(c)}(T) + \sum_{\ell\in\mathcal{K}\setminus\mathcal{C}} \Delta_\ell U_\ell^{(nc)}(T), \tag{17}$$

$$= (C-1)\cdot \mathrm{O}(\log T) + \mathrm{O}(1), \tag{18}$$

Here, $U_k^{(c)}(T)$ is the upper bound on $\mathbb{E}\left[n_k(T)\right]$ given in Theorem 8 and $U_\ell^{(nc)}(T)$ is the upper bound on $\mathbb{E}\left[n_\ell(T)\right]$ given in Theorem 9. The set $\mathcal{C}$ denotes the set of competitive arms and C is the cardinality of that set.

Observe that this regret bound is similar to the regret bound for UCB-C algorithm and enjoys the same properties of achieving bounded regret whenever possible and performing dimension reduction.

C 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 σ^2 sub-gaussian with mean μ, then*

$$\Pr(\mid \hat{\mu} - \mu \mid) \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2 T}{2\sigma^2}\right),$$

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

Lemma 1. *Probability that true mean lies outside the confidence set decays with total number of pulls,i.e t, as,*

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

Proof. See that,

$$\Pr\left(|\mu_k(\theta^*) - \hat{\mu}_{k,n_k(t)}| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}\right) \leq$$

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

$$\leq \sum_{m=1}^{t} 2t^{-\alpha} \quad (20)$$

$$= 2t^{1-\alpha}. \quad (21)$$

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. We use this trick repeatedly in the paper, whenever we encounter such expressions. 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 ϵ in Fact 1 being equal to $\sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}$. □

Lemma 2. *Define $E_1(t)$ to be the event that arm k^* is Θ_t-non-competitive for the round $t+1$, then,*

$$\Pr(E_1(t)) \leq 2t^{1-\alpha}.$$

Proof. Observe that,

$$\Pr(E_1(t)) \leq \Pr(\theta^* \notin \Theta_t) \quad (22)$$

$$= \Pr\left(|\mu_{k^{\max}}(\theta^*) - \hat{\mu}_{k^{\max}}| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_{k^{\max}}(t)}}\right) \quad (23)$$

$$\leq 2t^{1-\alpha} \quad (24)$$

Here (23) follows from definition of confidence set and (24) follows from Lemma 1. □

Lemma 3. *If $\Delta_{min} \geq 4\sqrt{\frac{K\alpha\sigma^2 \log t_0}{t_0}}$ for some constant $t_0 > 0$, then,*

$$\Pr(k_{t+1} = k | n_k(t) \geq s) \leq 6t^{1-\alpha} \quad \text{for } s \geq \frac{t}{2K}, \forall t > t_0,$$

where $k \neq k^*$ is a suboptimal arm.

Proof. The probability that arm k is pulled at step $t+1$, given it has been pulled s times can be bounded as follows:

$$\Pr(k_{t+1} = k | n_k(t) \geq s)$$

$$= \Pr(I_k(t) = \max_{k' \in \mathcal{C}_t} I_{k'}(t) | n_k(t) \geq s) \quad (25)$$

$$\leq \Pr(E_1(t) \cup (E_1^c(t), I_k(t) > I_{k^*}(t)) | n_k(t) \geq s) \quad (26)$$

$$\leq \Pr(E_1(t) | n_k(t) \geq s) + \Pr(E_1^c(t), I_k(t) > I_{k^*}(t) \mid n_k \geq s) \quad (27)$$

$$\leq 2t^{1-\alpha} + \Pr\left(I_k(t) > I_{k^*}(t) \mid n_k(t) \geq s\right). \quad (28)$$

Here, (27) follows from union bound and (28) follows from Lemma 2. We now bound the second term as,

$$\Pr(I_k(t) > I_{k^*}(t) | n_k(t) \geq s)$$

$$= \Pr\left(I_k(t) > I_{k^*}(t) | \mu_{k^*}(t) \leq I_{k^*}(t), n_k(t) \geq s\right) \times \Pr\left(\mu_{k^*}(t) \leq I_{k^*}(t) | n_k(t) \geq s\right) + \Pr\left(I_k(t) > I_{k^*}(t) | \mu_{k^*}(t) > I_{k^*}(t), n_k(t) \geq s\right) \times \Pr\left(\mu_{k^*}(t) > I_{k^*}(t) | n_k(t) \geq s\right) \quad (29)$$

$$\leq \Pr\left(I_k(t) > I_{k^*}(t) | \mu_{k^*}(t) \leq I_{k^*}(t), n_k(t) \geq s\right) + \Pr\left(\mu_{k^*}(t) > I_{k^*}(t) | n_k(t) \geq s\right) \quad (30)$$

$$\leq \Pr\left(I_k(t) > I_{k^*}(t) | \mu_{k^*}(t) \leq I_{k^*}(t), n_k(t) \geq s\right) + 2t^{1-\alpha} \quad (31)$$

$$= \Pr\left(I_k(t) > \mu_{k^*}(t) | n_k(t) \geq s\right) + 2t^{1-\alpha} \quad (32)$$

$$= \Pr\left(\hat{\mu}_k + \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}} > \mu_{k^*}(\theta^*) \middle| n_k(t) \geq s\right) + 2t^{1-\alpha} \quad (33)$$

$$= \Pr\left(\hat{\mu}_k - \mu_k(\theta^*) > \Delta_k - \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}} \middle| n_k(t) \geq s\right) + 2t^{1-\alpha} \quad (34)$$

$$\leq 2t \exp\left(-2s\left(\Delta_k - \sqrt{\frac{2\alpha\sigma^2 \log t}{s}}\right)^2\right) + 2t^{1-\alpha} \quad (35)$$

$$\leq 2t^{1-\alpha} \exp\left(-2s\left(\Delta_k^2 - 2\Delta_k\sqrt{\frac{2\alpha\sigma^2 \log t}{s}}\right)\right) + 2t^{1-\alpha} \quad (36)$$

$$= 4t^{1-\alpha} \quad \text{for all } t > t_0. \quad (37)$$

Equation (29) follows from the fact that $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$. Inequality (30) arrives from dropping $P(B)$ and $P(A|B^c)$ in the previous expression. We have (31) from Lemma 2 and the fact that $I_k(t) = \hat{\mu}_k + \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}$. Inequality (35) follows from the Hoeffding's inequality and the term t before the exponent in (35) arises as the random variable,$n_k(t)$, can take values between s and t. Equation (37) results from the definition of t_0 and the fact that $s > \frac{t}{2K}$.

Plugging the result of (37) in the expression (28) completes the proof of Lemma 3.

□

Lemma 4. *Consider a suboptimal arm $k \neq k^*$, which is ϵ_k-non-competitive. If $\epsilon_k \geq \sqrt{\frac{8\alpha\sigma^2 K \log t_0}{t_0}}$ for some constant $t_0 > 0$, then,*

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

Proof. We now bound this probability as,

$$\begin{aligned}
&\Pr(k_{t+1} = k | k^* = k^{\max}) \\
&= \Pr(k \in \mathcal{C}_t, I_k = \max_{\ell \in \mathcal{C}} I_\ell | k^* = k^{\max}) && (38)\\
&\leq \Pr(k \in \mathcal{C}_t | k^* = k^{\max}) && (39)\\
&\leq \Pr(|\hat{\mu}_{k^*} - \mu_{k^*}(\theta^*)| > \frac{\epsilon_k}{2} | k^* = k^{\max}) && (40)\\
&\leq 2t \exp\left(-\frac{\epsilon^2 t}{8K\sigma^2}\right) && (41)\\
&\leq 2t^{1-\alpha} \quad \forall t > t_0. && (42)
\end{aligned}$$

See that $|\hat{\mu}_{k^{\max}} - \mu_{k^{\max}}| < \frac{\epsilon_k}{2} \Rightarrow |\mu_{k^{\max}}(\theta) - \mu_{k^{\max}}(\theta^*)| < \epsilon$ for $\theta \in \tilde{\Theta}_t$. This holds as $\sqrt{\frac{2\alpha\sigma^2 \log t_0}{t_0}} \leq \frac{\epsilon_k}{2}$ and if $\theta \in \tilde{\Theta}_t$, then $|\mu_{k^{\max}}(\theta) - \hat{\mu}_{k^{\max}}| \leq \sqrt{\frac{2\alpha\sigma^2 \log t_0}{t_0}} \leq \frac{\epsilon_k}{2}$. Therefore in order for arm k to be $\tilde{\Theta}_t$-competitive, we need at least $|\hat{\mu}_{k^*} - \mu_{k^*}(\theta^*)| > \epsilon_k/2$, which leads to (40) as arm k is ϵ_k non-competitive. Inequality (41) follows from Hoeffding's inequality. The term t before the exponent in (41) arises as the random variable n_{k^*} can take values from $\frac{t}{K}$ to t. □

Lemma 5. *If $\Delta_{min} \geq 4\sqrt{\frac{K\alpha\sigma^2 \log t_0}{t_0}}$ for some constant $t_0 > 0$, then,*

$$\Pr\left(n_k(t) > \frac{t}{K}\right) \leq 6K\left(\frac{t}{K}\right)^{2-\alpha} \quad \forall t > Kt_0, k \neq k^*.$$

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

$$\begin{aligned}
&\Pr\left(n_k(t) \geq \frac{t}{K}\right) = \\
&\left(\Pr\left(n_k(t) \geq \frac{t}{K} \mid n_k(t-1) \geq \frac{t}{K}\right) \times \right. \\
&\left. \Pr\left(n_k(t-1) \geq \frac{t}{K}\right)\right) + \\
&\left(\Pr\left(k_t = k \mid n_k(t-1) = \frac{t}{K} - 1\right) \times \right. \\
&\left. \Pr\left(n_k(t-1) = \frac{n}{K} - 1\right)\right) && (43)\\
&\leq \Pr\left(n_k(t-1) \geq \frac{t}{K}\right) + \\
&\quad \Pr\left(k_t = k \mid n_k(t-1) = \frac{t}{K} - 1\right) && (44)\\
&\leq \Pr\left(n_k(t-1) \geq \frac{t}{K}\right) + \\
&\quad 6(t-1)^{1-\alpha} \quad \forall (t-1) > t_0. && (45)
\end{aligned}$$

Here (45) follows from Lemma 3.

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

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

Now consider the summation

$$\begin{aligned}
&\sum_{\tau=\frac{t}{K}}^{t} \Pr\left(n_k(\tau) \geq \frac{t}{K}\right) - \Pr\left(n_k(\tau-1) \geq \frac{t}{K}\right) \\
&\leq \sum_{\tau=\frac{t}{K}}^{t} 6(\tau-1)^{1-\alpha}. && (46)
\end{aligned}$$

This gives us,

$$\begin{aligned}
&\Pr\left(n_k(t) \geq \frac{t}{K}\right) - \Pr\left(n_k\left(\frac{t}{K} - 1\right) \geq \frac{t}{K}\right) \\
&\leq \sum_{\tau=\frac{t}{K}}^{t} 6(\tau-1)^{1-\alpha}. && (47)
\end{aligned}$$

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

$$\begin{aligned}
\Pr\left(n_k(t) \geq \frac{t}{K}\right) &\leq \sum_{\tau=\frac{t}{K}}^{t} 6(\tau-1)^{1-\alpha} && (48)\\
&\leq 6K\left(\frac{t}{K}\right)^{2-\alpha} \quad \forall t > Kt_0. && (49)
\end{aligned}$$

□

Proof of Theorem 2 We bound $\mathbb{E}[n_k(t)]$ as

$$\begin{aligned}
&\mathbb{E}[n_k(T)] = \\
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{k_t = k\}}\right] && (50)\\
&= \sum_{t=0}^{T-1} \Pr(k_{t+1} = k) && (51)\\
&= \sum_{t=1}^{Kt_0} \Pr(k_t = k) + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1} = k) && (52)
\end{aligned}$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Bigg(\Pr(n_{k^*}(t) = \max_{k'} n_{k'}(t)) \times \Pr(k_{t+1} = k | n_{k^*}(t) = \max_{k'} n_{k'}(t)) \Bigg) + \sum_{t=Kt_0}^{T-1} \sum_{k' \neq k^*} \Bigg(\Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \times \Pr(k_{t+1} = k | n_{k'}(t) = \max_{k''} n_{k''}(t)) \Bigg) \quad (53)$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1} = k | n_{k^*}(t) = \max_{k'} n_{k'}(t)) + \sum_{t=Kt_0}^{T-1} \sum_{k' \neq k^*} \Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \quad (54)$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} 2t^{1-\alpha} + \sum_{t=Kt_0}^{T} \sum_{k' \neq k^*} \Pr\left(n_{k'}(t) \geq \frac{t}{K} \right) \quad (55)$$

$$\leq Kt_0 + \sum_{t=1}^{T} 2t^{1-\alpha} + K(K-1) \sum_{t=Kt_0}^{T} 6 \left(\frac{t}{K} \right)^{2-\alpha}. \quad (56)$$

Here, (55) follows from Lemma 4 and (56) follows from Lemma 5.

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

$$\mathbb{E}[n_k(T)] \leq \sum_{t=1}^{T} \Pr(k_t = k) \quad (57)$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t) \cup (E_1^c(t), I_k > I_{k^*})) \quad (58)$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(E_1^c(t), I_k(t-1) > I_{k^*}(t-1)) \quad (59)$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(I_k(t-1) > I_{k^*}(t-1)) \quad (60)$$

$$= \sum_{t=1}^{T} 2t^{1-\alpha} + \sum_{t=0}^{T-1} \Pr(I_k(t) > I_{k^*}(t)) \quad (61)$$

$$= \sum_{t=1}^{T} 2t^{1-\alpha} + \mathbb{E}[\mathbb{1}_{I_k > I_{k^*}}(T)] \quad (62)$$

$$\leq 8\alpha\sigma^2 \frac{\log(T)}{\Delta_k^2} + \frac{2\alpha}{\alpha - 2} + \sum_{t=1}^{T} 2t^{1-\alpha}. \quad (63)$$

Here, (61) follows from Lemma 2. We have (63) from the analysis of UCB for the classical bandit problem, for details see proof of Theorem 2.1 in [23].

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

D Proofs for the UCB-int Algorithm

Lemma 6. *Define $E_1(t)$ to be the event that arm k^* is Θ_t-non-competitive for the round $t+1$, then,*

$$\Pr(E_1(t)) \leq 2Kt^{1-\alpha}.$$

Proof. Observe that,

$$\Pr(E_1(t)) \leq \Pr(\theta^* \notin \Theta_t) \quad (64)$$

$$= \Pr\left(\bigcup_{k \in \mathcal{K}} |\mu_k(\theta^*) - \hat{\mu}_{k,n_k(t)}| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}} \right) \quad (65)$$

$$\leq \sum_{k=1}^{K} \Pr\left(|\mu_k(\theta^*) - \hat{\mu}_{k,n_k(t)}| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}} \right) \quad (66)$$

$$\leq \sum_{k=1}^{K} \sum_{m=1}^{t} \Pr\left(|\mu_k(\theta^*) - \hat{\mu}_{k,m}| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{m}} \right) \quad (67)$$

$$\leq K \sum_{m=1}^{t} 2t^{-\alpha} \quad (68)$$

$$= 2Kt^{1-\alpha}. \quad (69)$$

We are using $\hat{\mu}_{k,m}$ to denote the empirical mean of rewards from arm k obtained from it's m pulls. Here (65) follows from definition of confidence set and (66) follows from union bound. We have (67) 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. Inequality (68) follows from Hoeffding's lemma. □

Lemma 7. *If $\Delta_{min} \geq 4\sqrt{\frac{K\alpha\sigma^2 \log t_0}{t_0}}$ for some constant $t_0 > 0$, then,*

$$\Pr(k_{t+1} = k | n_k(t) \geq s) \leq (2K+4)t^{1-\alpha} \quad \text{for } s \geq \frac{t}{2K},$$

$\forall t > t_0$, where $k \neq k^*$ is a suboptimal arm.

Proof. The probability that arm k is pulled at step $t+1$, given it has been pulled s times can be bounded as follows:

$$\Pr(k_{t+1} = k | n_k(t) \geq s)$$

$$= \Pr(I_k(t) = \max_{k' \in \mathcal{C}_t} I_{k'}(t) | n_k(t) \geq s) \quad (70)$$

$$\leq \Pr(E_1(t) \cup (E_1^c(t), I_k(t) > I_{k^*}(t)) | n_k(t) \geq s) \quad (71)$$

$$\leq \Pr(E_1(t)|n_k(t) \geq s) + \Pr(E_1^c(t), I_k(t) > I_{k^*}(t) \mid n_k \geq s) \quad (72)$$

$$\leq 2Kt^{1-\alpha} + \Pr\left(I_k(t) > I_{k^*}(t) \mid n_k(t) \geq s\right) \quad (73)$$

$$\leq (2K+4)t^{1-\alpha}. \quad (74)$$

We have (73) from Lemma 6. Second term is bounded by $4t^{-\alpha}$ from Lemma 3. □

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

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

Proof. The proof of Lemma 8 is the same as that of Lemma 4. □

Lemma 9. *If $\Delta_{min} \geq 4\sqrt{\frac{K\alpha\sigma^2 \log t_0}{t_0}}$ for some constant $t_0 > 0$, then,*

$$\Pr\left(n_k(t) > \frac{t}{K}\right) \leq 6K^2\left(\frac{t}{K}\right)^{2-\alpha} \quad \forall t > Kt_0, k \neq k^*.$$

Proof. The proof is very similar to that of Lemma 5 but with the difference that we have

$$\Pr(k_{t+1} = k | n_k(t) \geq s) \leq (2K+4)t^{1-\alpha} \quad \text{for } s \geq \frac{t}{2K},$$

$\forall t > t_0$ in this case.

As in proof of Lemma 5, we can write

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

$$\leq \Pr\left(n_k(t-1) \geq \frac{t}{K}\right) + \Pr\left(k_t = k \mid n_k(t-1) = \frac{t}{K} - 1\right) \quad (75)$$

$$\leq \Pr\left(n_k(t-1) \geq \frac{t}{K}\right) + 6K(t-1)^{1-\alpha} \quad \forall (t-1) > t_0. \quad (76)$$

Following the remaining steps of proof in Lemma 5, we arrive at the result that

$$\Pr\left(n_k(t) > \frac{t}{K}\right) \leq 6K^2\left(\frac{t}{K}\right)^{2-\alpha} \quad \forall t > Kt_0, k \neq k^*.$$

□

Proof of Theorem 6 We bound $\mathbb{E}[n_k(t)]$ as

$$\mathbb{E}[n_k(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{k_t=k\}}\right] \quad (77)$$

$$= \sum_{t=0}^{T-1} \Pr(k_{t+1} = k) \quad (78)$$

$$= \sum_{t=1}^{Kt_0} \Pr(k_t = k) + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1} = k) \quad (79)$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Bigg(\Pr(n_{k^*}(t) = \max_{k'} n_{k'}(t)) \times \Pr(k_{t+1} = k | n_{k^*}(t) = \max_{k'} n_{k'}(t)) \Bigg) + \sum_{t=Kt_0}^{T-1} \sum_{k' \neq k^*} \Bigg(\Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \times \Pr(k_{t+1} = k | n_{k'}(t) = \max_{k''} n_{k''}(t)) \Bigg) \quad (80)$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1} = k | n_{k^*}(t) = \max_{k'} n_{k'}(t)) + \sum_{t=Kt_0}^{T-1} \sum_{k' \neq k^*} \Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \quad (81)$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} 2Kt^{1-\alpha} + \sum_{t=Kt_0}^{T} \sum_{k' \neq k^*} \Pr\left(n_{k'}(t) \geq \frac{t}{K}\right) \quad (82)$$

$$\leq Kt_0 + \sum_{t=1}^{T} 2Kt^{1-\alpha} + K^2(K-1) \sum_{t=Kt_0}^{T} 6\left(\frac{t}{K}\right)^{2-\alpha}. \quad (83)$$

Here, (82) follows from Lemma 8 and (83) follows from Lemma 9.

Proof of Theorem 5 For any suboptimal arm $k \neq k^*$,

$$\mathbb{E}[n_k(T)] \leq \sum_{t=1}^{T} \Pr(k_t = k) \quad (84)$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t) \cup (E_1^c(t), I_k > I_{k^*})) \quad (85)$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(E_1^c(t), I_k(t-1) > I_{k^*}(t-1)) \quad (86)$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(I_k(t-1) > I_{k^*}(t-1)) \quad (87)$$

$$= \sum_{t=1}^{T} 2Kt^{1-\alpha} + \sum_{t=0}^{T-1} \Pr\left(I_k(t) > I_{k^*}(t)\right) \quad (88)$$

$$= \sum_{t=1}^{T} 2Kt^{1-\alpha} + \mathbb{E}\left[\mathbb{1}_{I_k > I_{k^*}}(T)\right] \quad (89)$$

$$\leq 8\alpha\sigma^2 \frac{\log(T)}{\Delta_k^2} + 2 + \sum_{t=1}^{T} \frac{2\alpha}{\alpha - 2} Kt^{1-\alpha}. \quad (90)$$

Here, (88) follows from Lemma 6. We have (90) from the analysis of UCB for the classical bandit problem, for details see proof of Theorem 2.1 in [23].

Proof of Theorem 7 Proof follows from the proof of Theorem 5 and Theorem 6.

E Proofs for the UCB-min Algorithm

Lemma 10. *Define $E_1(t)$ to be the event that arm k^* is Θ_t-non-competitive for the round $t+1$, then,*

$$\Pr(E_1(t)) \leq 2t^{1-\alpha}, \forall t > t_0$$

Proof. Observe that,

$$\Pr(E_1(t)) \leq \Pr(\theta^* \notin \Theta_t) \quad (91)$$

$$= \Pr\left(|\mu_{k^{\max}}(\theta^*) - \hat{\mu}_{k^{\max}}| \geq \frac{\delta}{2}\right) \quad (92)$$

$$\leq 2t \exp\left(-\frac{\delta^2 t}{8K\sigma^2}\right) \quad (93)$$

$$\leq 2t^{1-\alpha} \quad (94)$$

Observe that $|\mu_{k^{\max}}(\theta^*) - \hat{\mu}_{k^{\max}}| \leq \frac{\delta}{2} \Rightarrow |\mu_{k^{\max}}(\theta^*) - \mu_{k^{\max}}(\theta)| \leq \delta$, for all $\theta \in \Theta_t$. In order for arm k^* to non-competitive, we need $|\mu_{k^{\max}}(\theta) - \mu_{k^{\max}}(\theta^*)| > \delta$ (see assumption A1). This observation yields us inequality (92). Inequality (93) then follows from Hoeffding's inequality. The term t before the exponent in (93) arises as the random variable $n_{k^{\max}}$ can take values from 1 to t. □

Lemma 11. *The probability of selecting a suboptimal arm which has been played $\frac{t}{K}$ times is bounded as,*

$$\Pr(k_{t+1} = k | n_k(t) \geq s) \leq 6t^{1-\alpha} \quad \text{for } s \geq \frac{t}{2K}, \forall t > t_0,$$

where $k \neq k^*$ is a suboptimal arm.

Proof. The probability that arm k is pulled at step $t+1$, given it has been pulled s times can be bounded as follows:

$$\Pr(k_{t+1} = k | n_k(t) \geq s)$$

$$= \Pr(I_k(t) = \max_{k' \in \mathcal{C}_t} I_{k'}(t) | n_k(t) \geq s) \quad (95)$$

$$\leq \Pr(E_1(t) \cup (E_1^c(t), I_k(t) > I_{k^*}(t)) | n_k(t) \geq s) \quad (96)$$

$$\leq \Pr(E_1(t) | n_k(t) \geq s) + \Pr(E_1^c(t), I_k(t) > I_{k^*}(t) \mid n_k \geq s) \quad (97)$$

$$\leq 2t^{1-\alpha} + \Pr\left(I_k(t) > I_{k^*}(t) \mid n_k(t) \geq s\right) \quad (98)$$

$$\leq 6t^{1-\alpha}. \quad (99)$$

We have (98) from Lemma 10. Second term is bounded by $4t^{1-\alpha}$ from Lemma 3. □

Lemma 12. *Consider a suboptimal arm $k \neq k^*$, which is ϵ_k- non-competitive. If $\epsilon_k \geq \sqrt{\frac{8\alpha\sigma^2 K \log t_0}{t_0}}$ for some constant $t_0 > 0$, then,*

$$\Pr(k_{t+1} = k | k^* = k^{max}) \leq 2t \exp\left(-\frac{\epsilon_k^2 t}{8K\sigma^2}\right).$$

Proof. We now bound this probability as,

$$\Pr(k_{t+1} = k | k^* = k^{\max})$$

$$= \Pr(k \in \mathcal{C}_t, I_k = \max_{\ell \in \mathcal{C}} I_\ell | k^* = k^{\max}) \quad (100)$$

$$\leq \Pr(k \in \mathcal{C}_t | k^* = k^{\max}) \quad (101)$$

$$\leq \Pr\left(|\hat{\mu}_{k^*} - \mu_{k^*}(\theta^*)| > \frac{\epsilon_k}{2}\right) \quad (102)$$

$$\leq 2t \exp\left(-\frac{\epsilon_k^2 t}{8K\sigma^2}\right). \quad (103)$$

Notice that $|\hat{\mu}_{k^{\max}} - \mu_{k^{\max}}(\theta^*)| < \frac{\epsilon_k}{2} \Rightarrow |\mu_{k^{\max}}(\theta) - \mu_{k^{\max}}(\theta^*)| < \epsilon_k$ for $\theta \in \Theta_k$. Therefore, in order for arm k to be competitive, we need $|\hat{\mu}_{k^*} - \mu_{k^*}(\theta^*)| > \frac{\epsilon_k}{2}$ as arm k is ϵ_k non-competitive. This observation gives us inequality (102). Inequality (103) then follows from Hoeffding's inequality. The term t before the exponent in (103) arises as the random variable n_{k^*} can take values from $\frac{t}{K}$ to t. □

Lemma 13. *For the specified t_0,*

$$\Pr\left(n_k(t) > \frac{t}{K}\right) \leq 6K\left(\frac{t}{K}\right)^{2-\alpha} \quad \forall t > Kt_0, k \neq k^*.$$

Proof. Proof is the same as that of Lemma 5, with t_0 defined as

$$\inf\left\{\tau : \Delta_{\min} \geq 4\sqrt{\frac{K\alpha\sigma^2 \log \tau}{\tau}}, \delta \geq \sqrt{\frac{8K\alpha\sigma^2 \log \tau}{\tau}}\right\}.$$

□

Proof of Theorem 9 We bound $\mathbb{E}[n_k(t)]$ as

$$\begin{aligned}
&\mathbb{E}[n_k(T)] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{k_t=k\}}\right] && (104)\\
&= \sum_{t=0}^{T-1} \Pr(k_{t+1}=k) && (105)\\
&= \sum_{t=1}^{Kt_0} \Pr(k_t=k) + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1}=k) && (106)\\
&\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Bigg(\Pr(n_{k^*}(t) = \max_{k'} n_{k'}(t)) \times \\
&\quad \Pr(k_{t+1}=k | n_{k^*}(t) = \max_{k'} n_{k'}(t)) \Bigg) + \\
&\quad \sum_{t=Kt_0}^{T-1} \sum_{k' \neq k^*} \Bigg(\Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \times \\
&\quad \Pr(k_{t+1}=k | n_{k'}(t) = \max_{k''} n_{k''}(t)) \Bigg) && (107)\\
&\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1}=k | n_{k^*}(t) = \max_{k'} n_{k'}(t)) + \\
&\quad \sum_{t=Kt_0}^{T-1} \sum_{k' \neq k^*} \Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) && (108)\\
&\leq Kt_0 + \sum_{t=Kt_0}^{T-1} 2t \exp\left(-\frac{\epsilon_k^2 t}{8K\sigma^2}\right) + \\
&\quad \sum_{t=Kt_0}^{T} \sum_{k' \neq k^*} \Pr\left(n_{k'}(t) \geq \frac{t}{K}\right) && (109)\\
&\leq Kt_0 + \sum_{t=1}^{T} 2t \exp\left(-\frac{\epsilon_k^2 t}{8K\sigma^2}\right) + \\
&\quad K(K-1) \sum_{t=Kt_0}^{T} 6\left(\frac{t}{K}\right)^{2-\alpha}. && (110)
\end{aligned}$$

Here (109) follows from Lemma 12 and (110) follows from Lemma 13.

Proof of Theorem 8 For any suboptimal arm $k \neq k^*$,

$$\begin{aligned}
\mathbb{E}[n_k(T)] &\leq \sum_{t=1}^{T} \Pr(k_t=k) && (111)\\
&\leq \sum_{t=1}^{T} \Pr(E_1(t) \cup (E_1^c(t), I_k > I_{k^*})) && (112)\\
&\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \\
&\quad \Pr(E_1^c(t), I_k(t-1) > I_{k^*}(t-1)) && (113)\\
&\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(I_k(t-1) > I_{k^*}(t-1)) && (114)\\
&= \sum_{t=1}^{T} 2t \exp\left(-\frac{(\delta)^2 t}{8K\sigma^2}\right) + \\
&\quad \sum_{t=0}^{T-1} \Pr\left(I_k(t) > I_{k^*}(t)\right) && (115)\\
&= \sum_{t=1}^{T} 2t \exp\left(-\frac{(\delta)^2 t}{8K\sigma^2}\right) + \mathbb{E}\left[\mathbb{1}_{I_k > I_{k^*}}(T)\right] && (116)\\
&\leq 8\alpha\sigma^2 \frac{\log(T)}{\Delta_k^2} + \frac{2\alpha}{\alpha-2} \\
&\quad + \sum_{t=1}^{T} 2t \exp\left(-\frac{\delta^2 t}{8K\sigma^2}\right). && (117)
\end{aligned}$$

Here (115) follows from Lemma 10. We have (63) from the analysis of UCB for the classical bandit problem, for details see proof of Theorem 2.1 in [23].

Proof of Theorem 10: Follows directly by combining the results on Theorem 8 and Theorem 9.