Multi-Armed Bandits with Correlated Arms

Samarth Gupta, Shreyas Chaudhari, Gauri Joshi, Osman Yağan
Carnegie Mellon University

***Abstract*—We consider a multi-armed bandit framework where the rewards obtained by pulling different arms are correlated. We develop a unified approach to leverage these reward correlations and present fundamental generalizations of classic bandit algorithms to the correlated setting. We present a unified proof technique to analyze the proposed algorithms. Rigorous analysis of C-UCB (the correlated bandit versions of Upper-confidence-bound) reveals that the algorithm end up pulling certain sub-optimal arms, termed as *non-competitive*, only $\mathrm{O}(1)$ times, as opposed to the $\mathrm{O}(\log T)$ pulls required by classic bandit algorithms such as UCB, TS etc. We present regret-lower bound and show that when arms are correlated through a latent random source, our algorithms obtain *order-optimal* regret. We validate the proposed algorithms via experiments on the MovieLens and Goodreads datasets, and show significant improvement over classical bandit algorithms.**

Keywords: Multi-Armed Bandits, Online Learning, Sequential Decision Making, Regret Analysis

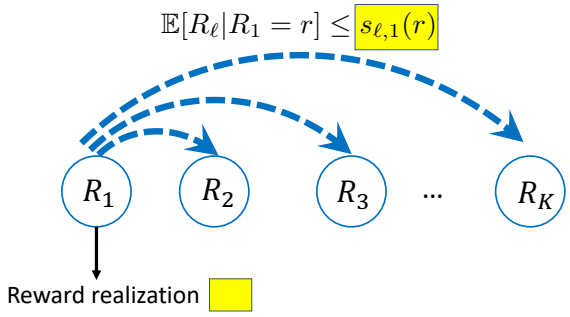

Fig. 1: Upon observing a reward r from an arm k, pseudo-rewards $s_{\ell,k}(r)$, give us an upper bound on the conditional expectation of the reward from arm ℓ given that we observed reward r from arm k. These pseudo-rewards model the correlation in rewards corresponding to different arms.

I. Introduction

A. Background and Motivation

Classical Multi-armed Bandits. The *multi-armed bandit* (MAB) problem falls under the class of sequential decision making problems. In the classical multi-armed bandit problem, there are K arms, with each arm having an *unknown* reward distribution. At each round t, we need to decide an arm $k_t \in \mathcal{K}$ and we receive a random reward R_{k_t} drawn from the reward distribution of arm k_t. The goal in the classical multi-armed bandit is to maximize the *long-term* cumulative reward. In order to maximize cumulative reward, it is important to balance the *exploration-exploitation* trade-off, i.e., pulling each arm enough number of times to identify the one with the highest mean reward, while trying to make sure that the arm with the highest mean reward is played as many times as possible. This problem has been well studied starting with the work of Lai and Robbins [1] that proposed the upper confidence bound (UCB) arm-selection algorithm and studied its fundamental limits in terms of bounds on *regret*. Subsequently, several other algorithms including Thompson Sampling (TS) [2] and KL-UCB [3], have been proposed for this setting. The generality of the classical multi-armed bandit model allows it to be useful in numerous applications. For example, MAB algorithms are useful in medical diagnosis [4], where the arms correspond to the different treatment mechanisms/drugs and are widely used for the problem of ad optimization [5] by viewing different version of ads as the arms in the MAB problem. The MAB framework is also useful in system testing [6], scheduling in computing systems [7], [8], [9], and web optimization [10], [5] .

Correlated Multi-Armed Bandits. The classical MAB setting implicitly assumes that the rewards are independent across arms, i.e., pulling an arm k does not provide any information about the reward we would have received from arm ℓ. However, this may not be true in practice as the reward corresponding to different treatment/drugs/ad-versions are likely to be *correlated* with each other. For instance, similar ads/drugs may generate similar reward for the user/patient. These correlations, when modeled and accounted for, can allow us to significantly improve the cumulative reward by reducing the amount of *exploration* in bandit algorithms.

Motivated by this, we study a variant of the classical multi-armed bandit problem in which rewards corresponding to different arms are correlated to each other, i.e., the conditional reward distribution satisfies $f_{R_\ell|R_k}(r_\ell|r_k) \neq f_{R_\ell}(r_\ell)$, whence $\mathbb{E}[R_\ell|R_k] \neq \mathbb{E}[R_\ell]$. Such correlations can only be learned upon obtaining samples from different arms simultaneously, i.e., by pulling multiple arms at a time. As that is not allowed in the classical Multi-Armed Bandit formulation, we assume the knowledge of such correlations in the form of prior knowledge that might be obtained through domain expertise or from controlled surveys. One way of capturing correlations is through the knowledge of the joint reward distribution. However, if the complete joint reward distribution is known, then the best-arm is known trivially. Instead, in our work, we only assume restrictive information about correlations in the form of *pseudo-rewards* that constitute an upper bound on conditional expected

[0] Email ids: samarthg@andrew.cmu.edu, schaudh2@andrew.cmu.edu, gaurij@andrew.cmu.edu, oyagan@andrew.cmu.edu. This paper was presented in part at the ICML Workshop on Theoretical Foundations of Reinforcement Learning 2020.

rewards. This makes our model more general and suitable for practical applications.

Fig. 1 presents an illustration of our model, where the pseudo-rewards, denoted by $s_{\ell,k}(r)$, provide an upper bound on the reward that we could have received from arm ℓ given that pulling arm k led to a reward of r; i.e.,

$$\mathbb{E}[R_\ell | R_k = r] \leq s_{\ell,k}(r). \quad (1)$$

We show that the knowledge of such bounds, even when they are not all tight, can lead to significant improvement in the cumulative reward obtained by reducing the amount of *exploration* compared to classical MAB algorithms. Our proposed MAB model and algorithm can be applied in all real-world applications of the classical Multi-Armed bandit problem, where it is possible to know pseudo-rewards from domain knowledge or through surveyed data. In the next section, we illustrate the applicability of our novel correlated Multi-Armed Bandit model and its differences with the existing contextual and structured bandit works through the example of optimal *ad-selection*.

B. An Illustrative Example

Suppose that a company is to run a display advertising campaign for one of their products, and its creative team have designed several different versions that can be displayed. It is expected that the user engagement (in terms of click probability and time spent looking at the ad) depends the version of the ad that is displayed. In order to maximize the total user engagement over the course of the ad campaign, multi-armed bandit algorithms can be used; different versions of the ad correspond to the *arms* and the reward from selecting an arm is given by the clicks or time spent looking at the ad version corresponding to that arm.

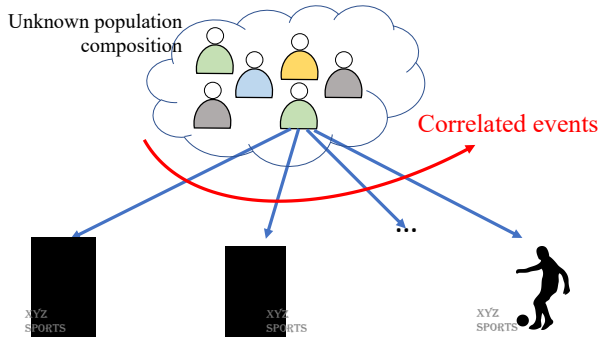

Fig. 2: The ratings of a user corresponding to different versions of the same ad are likely to be correlated. For example, if a person likes first version, there is a good chance that they will also like the 2nd one as it also related to tennis. However, the population composition is unknown, i.e., the fraction of people liking the first/second or the last version is unknown.

Personalized recommendations using Contextual and Structured bandits. Although the ad-selection problem can be solved by standard MAB algorithms, there are several specialized MAB variants that are designed to give better performance. For instance, the *contextual* bandit problem [11], [12] has been studied to provide *personalized* displays of the ads to the users. Here, before making a choice at each time step (i.e., deciding which version to show to a user), we observe the *context* associated with that user (e.g., age/occupation/income features). Contextual bandit algorithms learn the mappings from the context θ to the most favored version of ad $k^*(\theta)$ in an online manner and thus are useful for personalized recommendations. A closely related problem is the structured bandit problem [13], [14], [15], [16], in which the context θ (age/ income/ occupational features) is *hidden* but the mean rewards for different versions of ad (arms) as a function of hidden context θ are known. Such models prove useful for personalized recommendation in which the context of the user is unknown, but the reward mappings $\mu_k(\theta)$ are known through surveyed data.

Global Recommendations using Correlated-Reward Bandits. In this work we study a variant of the classical multi-armed bandit problem in which rewards corresponding to different arms are correlated to each other. In many practical settings, the reward we get from different arms at any given step are likely to be correlated. In the ad-selection example given in Figure 2, a user reacting positively (by clicking, ordering, etc.) to the first version of the ad with a girl playing tennis might also be more likely to click the second version as it is also related to tennis; of course one can construct examples where there is negative correlation between click events to different ads. The model we study in this paper explicitly captures these correlations through the knowledge of pseudo-rewards $s_{\ell,k}(r)$ (See Figure 1). Similar to the classical MAB setting, the goal here is to display versions of the ad to maximize user engagement. In addition, unlike contextual bandits, we do not observe the context (age/occupational/income) features of the user and do not focus on providing personalized recommendation. Instead our goal is to provide global recommendations to a population whose demographics is unknown. Unlike *structured bandits*, we do not assume that the mean rewards are functions of a hidden context parameter θ. In structured bandits, although the *mean* rewards depend on θ the reward realizations can still be independent. See Section II-D for more details.

C. Main Contributions and Organization

i) A General and Previously Unexplored Correlated Multi-Armed Bandit Model. In Section II we describe our novel correlated multi-armed bandit model, in which rewards of a user corresponding to different arms are correlated with each other. This correlation is captured by the knowledge of *pseudo-rewards*, which are upper bounds on the conditional mean reward of arm ℓ given reward of arm k. In practice, pseudo-rewards can be obtained via expert/domain knowledge (for example, common ingredients in two drugs that are being considered to treat an ailment) or controlled surveys (for example, beta-testing users who are asked to rate different versions of an ad). A key advantage of our framework is that pseudo-rewards are just upper bounds on the conditional expected rewards and can be arbitrarily loose. This also makes the proposed framework and algorithm directly usable in practice – if some pseudo-rewards are unknown due to lack

of domain knowledge/data, they can simply be replaced by the maximum possible reward entries, which serves a natural upper bound.

ii) An approach to generalize algorithms to the Correlated MAB setting. We propose a novel approach in Section III that extends any classical bandit (such as UCB, TS, KL-UCB etc.) algorithm to the correlated MAB setting studied in this paper. This is done by making use of the pseudo-rewards to reduce exploration in standard bandit algorithms. We refer to this algorithm as C-BANDIT where BANDIT refers to the classical bandit algorithm used in the last step of the algorithm (i.e., UCB/TS/KL-UCB).

iii) Unified Regret Analysis We study the performance of our proposed algorithms by analyzing their expected *regret*, $\mathbb{E}\left[\mathrm{Reg}(T)\right]$. The regret of an algorithm is defined as the difference between the cumulative reward of a *genie* policy, that always pulls the optimal arm k^*, and the cumulative reward obtained by the algorithm over T rounds. By doing regret analysis of C-UCB, we obtain the following upper bound on the expected regret of C-UCB.

Proposition 1 (Upper Bound on Expected Regret)**.** *The expected cumulative regret of the C-UCB algorithm is upper bounded as*

$$\mathbb{E}\left[Reg(T)\right] \leq (C-1)\cdot \mathrm{O}(\log T) + \mathrm{O}(1), \tag{2}$$

Here C denotes the number of *competitive* arms. An arm k is said to be *competitive* if expected pseudo-reward of arm k with respect to the optimal arm k^* is larger than the mean reward of arm k^*, that is, if $\mathbb{E}\left[s_{k,k^*}(r)\right] \geq \mu_{k^*}$, otherwise, the arm is said to be non-competitive. The result in Proposition 1 arises from the fact that the C-UCB algorithm ends up pulling the non-competitive arms only $\mathrm{O}(1)$ times and only the competitive sub-optimal arms are pulled $\mathrm{O}(\log T)$ times. In contrast to UCB, that pulls all $K-1$ sub-optimal arms $\mathrm{O}(\log T)$ times, our proposed C-UCB algorithm pulls only $C-1 \leq K-1$ arms $\mathrm{O}(\log T)$ times. In fact, when $C=1$, our proposed algorithm achieves *bounded* regret meaning that after some finite step, no arm but the optimal one will be selected. In this sense, we reduce a K-armed bandit problem to a C-armed bandit problem. We emphasize that k^*, μ^* and C are *all* unknown to the algorithm at the beginning.

We present our detailed regret bounds and analysis in Section IV. A rigorous analysis of the regret achieved under C-UCB is given through a unified technique. This technique can be of broad interest as we also provide a recipe to obtain regret analysis for any *C-Bandit* algorithm. For instance, the analysis of C-KL-UCB can be easily done through our provided outline.

iv) Evaluation using real-world datasets.

We perform simulations to validate our theoretical results in Section V. In Section VI, we do extensive validation of our results by performing experiments on two real-world datasets, namely MOVIELENS and GOODREADS, which show that the proposed approach yields drastically smaller regret than classical Multi-Armed Bandit strategies.

r	$s_{2,1}(r)$
0	0.7
1	0.4

r	$s_{1,2}(r)$
0	0.8
1	0.5

(a)	$R_1=0$	$R_1=1$
$R_2=0$	0.2	0.4
$R_2=1$	0.2	0.2

(b)	$R_1=0$	$R_1=1$
$R_2=0$	0.2	0.3
$R_2=1$	0.4	0.1

TABLE I: The top row shows the pseudo-rewards of arms 1 and 2, i.e., upper bounds on the conditional expected rewards (which are known to the player). The bottom row depicts two possible joint probability distribution (unknown to the player). Under distribution (a), Arm 1 is optimal whereas Arm 2 is optimal under distribution (b).

II. PROBLEM FORMULATION

A. Correlated Multi-Armed Bandit Model

Consider a Multi-Armed Bandit setting with K arms $\{1,2,\dots K\}$. At each round t, a user enters the system and we need to decide an arm k_t to display to the user. Upon pulling arm k_t, we receive a random reward $R_{k_t} \in [0,B]$. Our goal is to maximize the cumulative reward over time. The expected reward of arm k, is denoted by μ_k. If we knew the arm with highest mean, i.e., $k^* = \arg\max_{k\in\mathcal{K}} \mu_k$ beforehand, then we would always pull arm k^* to maximize expected cumulative reward. We now define the cumulative regret, minimizing which is equivalent to maximizing cumulative reward:

$$Reg(T) = \sum_{t=1}^{T} \mu_{k_t} - \mu_{k^*} = \sum_{k\neq k^*} n_k(T)\Delta_k. \tag{3}$$

Here, $n_k(T)$ denotes the number of times a sub-optimal arm is pulled till round T and Δ_k denotes the *sub-optimality gap* of arm k, i.e., $\Delta_k = \mu_{k^*} - \mu_k$.

The classical multi-Armed bandit setting implicitly assumes the rewards $R_1, R_2 \dots R_K$ are independent, that is, $\Pr(R_\ell = r_\ell | R_k = r) = \Pr(R_\ell = r_\ell) \quad \forall r_\ell, r \& \forall \ell, k$, which implies that, $\mathbb{E}\left[R_\ell | R_k = r\right] = \mathbb{E}\left[R_\ell\right] \quad \forall r, \ell, k$. However, in most practical scenarios this assumption is unlikely to be true. In fact, rewards of a user corresponding to different arms are likely to be correlated. Motivated by this we consider a setup where the conditional distribution of the reward from arm ℓ given reward from arm k is not equal to the probability distribution of the reward from arm ℓ, i.e., $f_{R_\ell|R_k}(r_\ell|r_k) \neq f_{R_\ell}(r_\ell)$, with $f_{R_\ell}(r_\ell)$ denoting the probability distribution function of the reward from arm ℓ. Consequently, due to such correlations, we have $\mathbb{E}\left[R_\ell|R_k\right] \neq \mathbb{E}\left[R_\ell\right]$.

In our problem setting, we consider that the player has partial knowledge about the joint distribution of correlated arms in the form of *pseudo-rewards*, as defined below:

Definition 1 (Pseudo-Reward)**.** *Suppose we pull arm k and observe reward r, then the pseudo-reward of arm ℓ with respect to arm k, denoted by $s_{\ell,k}(r)$, is an upper bound on the conditional expected reward of arm ℓ, i.e.,*

$$\mathbb{E}[R_\ell | R_k = r] \leq s_{\ell,k}(r), \tag{4}$$

without loss of generality, we define $s_{\ell,\ell}(r) = r$.

The pseudo-rewards information consists of a set of $K \times K$ functions $s_{\ell,k}(r)$ over $[0,B]$. This information can be obtained

r	$s_{2,1}(r)$	$s_{3,1}(r)$
0	0.7	2
1	0.8	1.2
2	2	1

r	$s_{1,2}(r)$	$s_{3,2}(r)$
0	0.5	1.5
1	1.3	2
2	2	0.8

r	$s_{1,3}(r)$	$s_{2,3}(r)$
0	1.5	2
1	2	1.3
2	0.7	0.75

TABLE II: If some pseudo-reward entries are unknown (due to lack of prior-knowledge/data), those entries can be replaced with the maximum possible reward and then used in the C-BANDIT algorithm. We do that here by entering 2 for the entries where pseudo-rewards are unknown.

in practice through either domain/expert knowledge or from controlled surveys. For instance, in the context of medical testing, where the goal is to identify the best drug to treat an ailment from among a set of K possible options, the effectiveness of two drugs is correlated when the drugs share some common ingredients. Through domain knowledge of doctors, it is possible answer questions such as "what are the chances that drug B would be effective given drug A was not effective?", through which we can infer the pseudo-rewards.

B. Computing Pseudo-Rewards from prior-data/surveys

The pseudo-rewards can also be learned from prior-available data, or through *offline* surveys in which users are presented with *all* K arms allowing us to sample $R_1, \ldots, R_K$ jointly. Through such data, we can evaluate an estimate on the conditional expected rewards. For example in Table I, we can look at all users who obtained 0 reward for Arm 1 and calculate their average reward for Arm 2, say $\hat{\mu}_{2,1}(0)$. This average provides an estimate on the conditional expected reward. Since we only need an upper bound on $\mathbb{E}[R_2|R_1 = 0]$, we can use several approaches to construct the pseudo-rewards.

1) If the training data is *large*, one can use the empirical estimate $\hat{\mu}_{2,1}(0)$ directly as $s_{2,1}(0)$, because through law of large numbers, the empirical average equals the $\mathbb{E}[R_2|R_1 = 0]$.
2) Alternatively, we can set $s_{2,1}(0) = \hat{\mu}_{2,1}(0) + \hat{\sigma}_{2,1}(0)$, with $\hat{\sigma}_{2,1}(0)$ denoting the empirical standard deviation on the conditional reward of arm 2, to ensure that pseudo-reward is an upper bound on the conditional expected reward.
3) In addition, pseudo-rewards for any unknown conditional mean reward could be filled with the maximum possible reward for the corresponding arm. Table II shows an example of a 3-armed bandit problem where some pseudo-reward entries are unknown, e.g., due to lack of data. We can fill these missing entries with maximum possible reward $(i.e., 2)$ as shown in Table II to complete the pseudo-reward entries.
4) If through the training data, we obtain a soft upper bound u on $\mathbb{E}[R_2|R_1 = 0]$ that holds with probability $1 - \delta$, then we can translate it to the pseudo-reward $s_{2,1}(0) = u \times (1 - \delta) + 2 \times \delta$, (assuming maximum possible reward is 2).

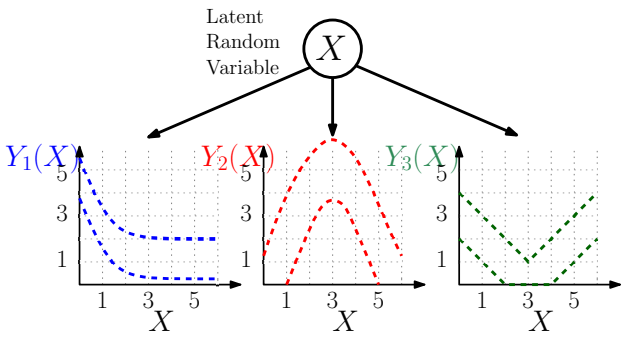

Fig. 3: A special case of our proposed problem framework is a setting in which rewards for different arms are correlated through a hidden random variable X. At each round X takes a realization in $\mathcal{X}$. The reward obtained from an arm k is $Y_k(X)$. The figure illustrates lower bounds and upper bounds on $Y_k(X)$ (through dotted lines). For instance, when X takes the realization 1, reward of arm 3 is a random variable bounded between 1 and 3.

Remark 1. *Note that the pseudo-rewards are upper bounds on the expected conditional reward and not hard bounds on the conditional reward itself. This makes our problem setup practical as upper bounds on expected conditional reward are easier to obtain, as illustrated in the previous paragraph.*

Remark 2 (Reduction to Classical Multi-Armed Bandits)**.** *When all pseudo-reward entries are unknown, then all pseudo-reward entries can be filled with maximum possible reward for each arm, that is,* $s_{\ell,k}(r) = B \ \forall r, \ell, k$. *In such a case, the problem framework studied in this paper reduces to the setting of the classical Multi-Armed Bandit problem and our proposed* C-BANDIT *algorithm performs exactly as standard* BANDIT *(for e.g., UCB, TS etc.) algorithms.*

While the pseudo-rewards are known in our setup, the underlying joint probability distribution of rewards is unknown. For instance, Table I (a) and Table I (b) show two joint probability distributions of the rewards that are both possible given the pseudo-rewards at the top of Table I. If the joint distribution is as given in Table I (a), then Arm 1 is optimal, while Arm 2 is optimal if the joint distribution is as given in Table I(b).

Remark 3. *For a setting where reward domain is* large *or there are a large number of arms, it may be difficult to learn the pseudo-reward entries from prior-data. In such scenarios, the knowledge of additional correlation structure may be helpful to know the value of pseudo-rewards. We describe one such structure in the next section where rewards are correlated through a latent random source and show how to evaluate pseudo-rewards in such a scenario.*

C. Special Case: Correlated Bandits with a Latent Random Source

Our proposed correlated multi-armed bandit framework subsumes many interesting and previously unexplored multi-armed bandit settings. One such special case is the correlated multi-armed bandit model where the rewards depend on a

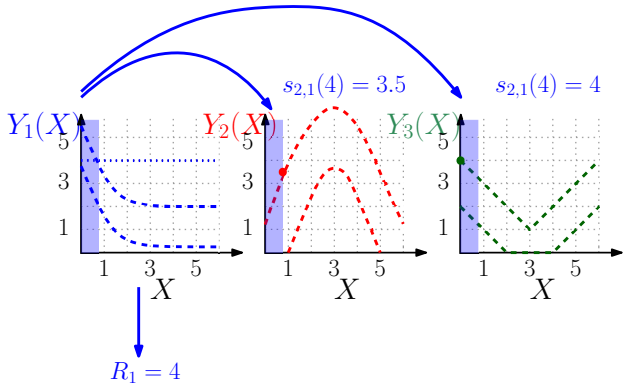

Fig. 4: An illustration on how to calculate pseudo-rewards in CMAB with latent random source. Upon observing a reward of 4 from arm 1, we can see that the maximum possible reward for arms 2 and 3 is 3.5 and 4 respectively. Therefore, $s_{2,1}(4) = 3.5$ and $s_{3,1}(4) = 4$.

common latent source of randomness [17]. More concretely, the rewards of different arms are correlated through a hidden random variable X (see Figure 3). At each round t, X takes a an i.i.d. realization $X_t \in \mathcal{X}$ (unobserved to the player) and upon pulling arm k, we observe a random reward $Y_k(X_t)$. The latent random variable X here could represent the *features* (i.e., age/occupation etc.) of the user arriving to the system, to whom we show one of the K arms. These *features* of the user are hidden in the problem due to privacy concerns. The random reward $Y_k(X_t)$ represents the preference of user with context X_t for the k^{th} version of the ad, for the application of ad-selection.

In this problem setup, upper and lower bounds on $Y_k(X)$, namely $\bar{g}_k(X)$ and $\underline{g}_k(X)$ are known. For instance, the information on upper and lower bounds of $Y_k(X_t)$ could represent knowledge of the form that *children of age 5-10 rate documentaries only in the range 1-3 out of 5*. Such information can be known or learned through prior available data. While the bounds on $Y_k(X)$ are known, the distribution of X and reward distribution within the bounds is unknown, due to which the optimal arm is not known beforehand. Thus, an online approach is needed to minimize the regret.

It is possible to translate this setting to the general framework described in the problem by transforming the mappings $Y_k(X)$ to pseudo-rewards $s_{\ell,k}(r)$. Recall the pseudo-rewards represent an upper bound on the conditional expectation of the rewards. In this framework, $s_{\ell,k}(r)$ can be calculated as:

$$s_{\ell,k}(r) = \max_{\{x:\underline{g}_k(x)\le r\le \bar{g}_k(x)\}} \bar{g}_\ell(x),$$

where $\underline{g}_k(x)$ and $\bar{g}_k(x)$ represent upper and lower bounds on $Y_k(x)$. Upon observing a realization from arm k, it is possible to estimate the maximum possible reward that would have been obtained from arm ℓ through the knowledge of bounds on $Y_k(X)$.

Figure 4 illustrates how pseudo-reward is evaluated when we obtain a reward $r = 4$ by pulling arm 1. We first infer that X lies in $[0, 0.8]$ if $r = 4$ and then find the maximum possible reward for arm 2 and arm 3 in $[0, 0.8]$. Once these pseudo-rewards are constructed, the problem fits in the general framework described in this paper and we can use the algorithms proposed for this setting directly.

Remark 4. *In the scenario where $\underline{g}_k(x)$ and $\bar{g}_k(x)$ are soft lower and upper bounds, i.e., $\underline{g}_k(x) \le Y_k(x) \le \bar{g}_k(x)$ w.p. $1-\delta$, we can still construct pseudo-reward as follows:*

$$s_{\ell,k}(r) = (1-\delta)^2 \times \left(\max_{\{x:\underline{g}_k(x)\le r\le \bar{g}_k(x)\}} \bar{g}_\ell(x) \right) + (1-(1-\delta)^2) \times M,$$

where M is the maximum possible reward an arm can provide. Thus our proposed framework and algorithms work under this setting as well.[1]

D. Comparison with parametric (structured) models

As mentioned in Section 1, a seemingly related model is the structured bandits model [13], [14], [18]. Structured bandits is a class of problems that cover linear bandits [15], generalized linear bandits [19], Lipschitz bandits [20], global bandits [**?**], regional bandits [21] etc. In the structured bandits setup, mean rewards corresponding to different arms are related to one another through a hidden parameter θ. The underlying value of θ is fixed and the mean reward mappings $\theta \to \mu_k(\theta)$ are known. Similarly, [22] studies a dependent armed bandit problem, that also has mean rewards corresponding to different arms related to one another. It considers a parametric model, where mean rewards of different arms are drawn from one of the K clusters, each having an unknown parameter π_i. All of these models are fundamentally different from the problem setting considered in this paper.

We list some of the differences with the structured bandits (and the model in [22]) below.

1) In this work we explicitly model the correlations in the rewards of a user corresponding to different arms. While mean rewards are related to each other in structured bandits and [22], the reward realizations are not necessarily correlated.
2) The model studied here is non-parametric in the sense that there is no hidden feature space as is the case in structured bandits and the work of Pandey et al. [22].
3) In structured bandits, the reward mappings from θ to $\mu_k(\theta)$ need to be *exact*. If they happen to be incorrect, then the algorithms for structured bandit cannot be used as they rely on the correctness of $\mu_k(\theta)$ to construct confidence intervals on the unknown parameter θ. In contrast, the model studied here relies on the pseudo-rewards being upper bounds on conditional expectations. These bounds need not be tight and the proposed C-Bandit algorithms adjust accordingly and perform at least as well as the corresponding classical bandit algorithm.
4) Similar to the structured bandits, the unimodal bandit framework [23], [24] also assumes a structure on the

[1]We evaluate a range of values within which x lies based on the reward with probability $1-\delta$. The maximum possible reward of arm ℓ for values of x is then identified with probability $1-\delta$. Due to this, with probability $(1-\delta)^2$, conditional reward of arm ℓ is at-most $\max_{\{x:\underline{g}_k(x)\le r\le \bar{g}_k(x)\}} \bar{g}_\ell(x)$.

mean rewards and does not capture the reward correlations explicitly. Under the unimodal framework, it is assumed that the mean reward μ_k as a function of the arms k has a single mode. Instead of assuming that mean rewards are related to one another, our framework explicitly captures the inherent correlations in the form of pseudo-reward. Unimodal bandits have often been used to model the problem of link-rate adaptation in wireless networks, where the mean-reward corresponding to different choices of arms is a unimodal function [25], [26], [27]. The same problem can also be dealt by modeling the correlations explicitly through the pseudo-reward framework described in this paper.

III. The Proposed C-Bandit Algorithms

We now propose an approach that extends the classical multi-armed bandit algorithms (such as UCB, Thompson Sampling, KL-UCB) to the correlated MAB setting. At each round $t+1$, the UCB algorithm [28] selects the arm with the highest UCB index $I_{k,t}$, i.e.,

$$k_{t+1} = \arg\max_{k\in\mathcal{K}} I_{k,t}, \quad I_{k,t} = \hat{\mu}_k(t) + B\sqrt{\frac{2\log(t)}{n_k(t)}}, \tag{5}$$

where $\hat{\mu}_k(t)$ is the empirical mean of the rewards received from arm k until round t, and $n_k(t)$ is the number of times arm k is pulled till round t. The second term in the UCB index causes the algorithm to explore arms that have been pulled only a few times (i.e., those with small $n_k(t)$). Recall that we assume all rewards to be bounded within an interval of size B. When the index t is implied by context, we abbreviate $\hat{\mu}_k(t)$ and $I_k(t)$ to $\hat{\mu}_k$ and I_k respectively in the rest of the paper.

Under Thompson sampling [29], the arm $k_{t+1} = \arg\max_{k\in\mathcal{K}} S_{k,t}$ is selected at time step $t+1$. Here, $S_{k,t}$ is the sample obtained from the posterior distribution of μ_k, That is,

$$k_{t+1} = \arg\max_{k\in\mathcal{K}} S_{k,t}, \quad S_{k,t} \sim \mathcal{N}\left(\hat{\mu}_k(t), \frac{\beta B}{n_k(t)+1}\right), \tag{6}$$

here β is a hyperparameter for the Thompson Sampling algorithm

In the correlated MAB framework, the rewards observed from one arm can help estimate the rewards from other arms. Our key idea is to use this information to reduce the amount of exploration required. We do so by evaluating the *empirical pseudo-reward* of every other arm ℓ with respect to an arm k at each round t. Using this additional information, we identify some arms as *empirically non-competitive* at round t, and only for this round, do not consider them as a candidate in the UCB/Thompson Sampling/(any other bandit algorithm).

A. Empirical Pseudo-Rewards

In our correlated MAB framework, pseudo-reward of arm ℓ with respect to arm k provides us an estimate on the reward of arm ℓ through the reward sample obtained from arm k. We now define the notion of empirical pseudo-reward which can be used to obtain an *optimistic estimate* of μ_ℓ through just reward samples of arm k.

Definition 2 (Empirical and Expected Pseudo-Reward)**.** *After t rounds, arm k is pulled $n_k(t)$ times. Using these $n_k(t)$ reward realizations, we can construct the empirical pseudo-reward $\hat{\phi}_{\ell,k}(t)$ for each arm ℓ with respect to arm k as follows.*

$$\hat{\phi}_{\ell,k}(t) \triangleq \frac{\sum_{\tau=1}^{t} \mathbb{1}_{k_\tau=k}\ s_{\ell,k}(r_{k_\tau})}{n_k(t)}, \qquad \ell \in \{1,\dots,K\}\setminus\{k\}, \tag{7}$$

The expected pseudo-reward of arm ℓ with respect to arm k is defined as

$$\phi_{\ell,k} \triangleq \mathbb{E}\left[s_{\ell,k}(R_k)\right]. \tag{8}$$

For convenience, we set $\hat{\phi}_{k,k}(t) = \hat{\mu}_k(t)$ and $\phi_{k,k} = \mu_k$.

Observe that $\mathbb{E}\left[s_{\ell,k}(R_k)\right] \geq \mathbb{E}\left[\mathbb{E}\left[R_\ell | R_k = r\right]\right] = \mu_\ell$. Due to this, empirical pseudo-reward $\hat{\phi}_{\ell,k}(t)$ can be used to obtain an estimated upper bound on μ_ℓ. Note that the empirical pseudo-reward $\hat{\phi}_{\ell,k}(t)$ is defined with respect to arm k and it is only a function of the rewards observed by pulling arm k.

B. The C-Bandit Algorithm

Using the notion of empirical pseudo-rewards, we now describe a 3-step procedure to fundamentally generalize classical bandit algorithms for the correlated MAB setting.

Step 1: Identify the set $\mathcal{S}_t$ of significant arms: At each round t, define $\mathcal{S}_t$ to be the set of arms that have at least t/K samples, i.e., $\mathcal{S}_t = \{k \in \mathcal{K} : n_k(t) > \frac{t}{K}\}$. As $\mathcal{S}_t$ is the set of arms that have relatively *large* number of samples, we use these arms for the purpose of identifying *empirically competitive* and *empirically non-competitive* arms. Furthermore, define $k^{\text{emp}}(t)$ to be the arm that has the highest empirical mean in set $\mathcal{S}_t$, i.e., $k^{\text{emp}}(t) = \arg\max_{k\in\mathcal{S}_t} \hat{\mu}_k(t)$. [2]

Step 2: Identify the set of *empirically competitive* arms $\mathcal{A}_t$:

Using the empirical mean, $\hat{\mu}_{k^{\text{emp}}}(t)$, of the arm with highest empirical reward in the set $\mathcal{S}_t$, we define the notions of empirically non-competitive and empirically competitive arms below.

Definition 3 (Empirically Non-Competitive arm at round t)**.** *An arm k is said to be Empirically Non-Competitive at round t, if $\min_{\ell\in\mathcal{S}_t} \hat{\phi}_{k,\ell}(t) < \hat{\mu}_{k^{emp}}(t)$.*

Definition 4 (Empirically Competitive arm at round t)**.** *An arm k is said to be Empirically Competitive at round t if $\min_{\ell\in\mathcal{S}_t} \hat{\phi}_{k,\ell}(t) \geq \hat{\mu}_{k^{emp}}(t)$. The set of all empirically competitive arms at round t is denoted by $\mathcal{A}_t$.*

The expression $\min_{\ell\in\mathcal{S}_t} \hat{\phi}_{k,\ell}(t)$ provides the tightest estimated upper bound on mean of arm k, through the samples of arms in $\mathcal{S}_t$. If this estimated upper bound is smaller than the estimated mean of $k^{\text{emp}}(t)$, then we call arm k as *empirically non-competitive* as it seems unlikely to be optimal through the samples of arms in $\mathcal{S}_t$. If the estimated upper bound of arm k

[2] If one were to use all arms (even those that have few samples) to identify empirically non-competitive arms, it can lead to incorrect inference, as pseudo-rewards with few samples will have larger noise, which can in-turn lead to elimination of the optimal arm. Using only the arms that have been pulled $\frac{t}{K}$ times in $\mathcal{S}_t$, allows us to ensure that the non-competitive arms are pulled only $\mathrm{O}(1)$ times as we show in Section IV.

is greater than $\hat{\mu}_{k^{\text{emp}}}(t)$, i.e., $\min_{\ell \in \mathcal{S}_t} \hat{\phi}_{k,\ell}(t) \geq \hat{\mu}_{k^{\text{emp}}}(t)$, we call arm k as empirically competitive at round t, as it cannot be inferred as sub-optimal through samples of arms in $\mathcal{S}_t$. Note that the set of empirically competitive and empirically non-competitive arms is evaluated at each round t and hence an arm that is empirically non-competitive at round t may be empirically competitive in subsequent rounds.

Step 3: Play BANDIT algorithm in $\{\mathcal{A}_t \cup \{k^{\textbf{emp}}(t)\}\}$ As empirically non-competitive arm seem sub-optimal to be selected at round t, we only consider the set of empirically competitive arms along with $k^{\text{emp}}(t)$ in this step of the algorithm. At round t, we play a BANDIT algorithm from the set $\mathcal{A}_t \cup \{k^{\text{emp}}(t)\}$. For instance, the C-UCB pulls the arm

$$k_t = \arg \max_{k \in \{\mathcal{A}_t \cup k^{emp}(t)\}} I_{k,t-1},$$

where $I_{k,t-1}$ is the UCB index defined in (5).

Similarly, C-TS pulls the arm

$$k_t = \arg \max_{k \in \{\mathcal{A}_t \cup k^{emp}(t)\}} S_{k,t-1},$$

where $S_{k,t}$ is the Thompson sample defined in (6)). At the end of each round we update the empirical pseudo-rewards $\hat{\phi}_{\ell,k_t}(t)$ for all ℓ, the empirical reward for arm k_t.

Note that our C-BANDIT approach allows using any classical Multi-Armed Bandit algorithm in the correlated Multi-Armed Bandit setting. This is important because some algorithms such as Thompson Sampling and KL-UCB are known to obtain much better empirical performance over UCB. Extending those to the correlated MAB setting allows us to have the superior empirical performance over UCB even in the correlated setting. This benefit is demonstrated in our simulations and experiments described in Section V and Section VI.

Remark 5 (Pseudo-lower bounds)**.** *If suppose one had the information about pseudo-lower bounds (which are lower bounds on conditional expected rewards), then it is possible to use this in our correlated bandit framework. In step 2 of our algorithm, we identify an arm k as empirically non-competitive if $\min_{\ell \in \mathcal{S}_t} \hat{\phi}_{k,\ell}(t) < \hat{\mu}_k^{emp}(t)$. We can maintain empirical pseudo-lower bound $\hat{w}_{i,j}(t)$ of each arm i with respect to every other arm j. Then, we can replace the step 2 of our algorithm by calling an arm empirically non-competitive if $\min_{\ell \in \mathcal{S}_t} \hat{\phi}_{k,\ell}(t) < \max_{i \in \mathcal{S}_t} \max_{j \in \mathcal{S}_t} \hat{w}_{i,j}(t)$. In the situation where pseudo-lower bounds are unknown, they can be set to $-\infty$ and the algorithm reduces to the C-Bandit algorithm proposed in the paper. We can expect the empirical performance of this algorithm (which is aware of pseudo-lower bounds) to be slightly better than the C-Bandit algorithm. However, its regret guarantees will be the same as that of the C-Bandit algorithm. This is because pseudo-upper bounds are crucial to deciding whether an arm is competitive/non-competitive (defined in the next section), and pseudo-lower bounds are not. Put differently, even in the presence of pseudo-lower bound, the definition of non-competitive and competitive arms (Definition 5) remains the same.*

Algorithm 1 C-UCB Correlated UCB Algorithm

Input: Pseudo-rewards $s_{\ell,k}(r)$
Initialize: $n_k = 0, I_k = \infty$ for all $k \in \{1, 2, \dots K\}$
for each round t **do**
  Find $\mathcal{S}_t = \{k : n_k(t) \geq \frac{t}{K}\}$, the arm that have been pulled significant number of times till $t-1$. Define $k^{\text{emp}}(t) = \arg\max_{k \in \mathcal{S}_t} \hat{\mu}_k(t)$.
  Initialize the empirically competitive set $\mathcal{A}_t$ as an empty set $\{\}$.
  for $k \in \mathcal{K}$ **do**
    if $\min_{\ell \in \mathcal{S}_t} \hat{\phi}_{k,\ell}(t) \geq \hat{\mu}_{k^{\text{emp}}}(t)$ **then**
      Add arm k to the empirically competitive set: $\mathcal{A}_t = \mathcal{A}_t \cup \{k\}$
    end if
  end for
  Apply UCB1 over arms in $\mathcal{A}_t \cup \{k^{\text{emp}}(t)\}$ by pulling arm $k_t = \arg\max_{k \in \mathcal{A}_t \cup \{k^{\text{emp}}(t)\}} I_k(t-1)$
  Receive reward r_t, and update $n_{k_t}(t) = n_{k_t}(t) + 1$
  Update Empirical reward: $\hat{\mu}_{k_t}(t) = \frac{\hat{\mu}_{k_t}(t-1)(n_{k_t}(t)-1)+r_t}{n_{k_t}(t)}$
  Update the UCB Index: $I_{k_t}(t) = \hat{\mu}_{k_t}(t) + B\sqrt{\frac{2\log t}{n_{k_t}(t)}}$
  Update empirical pseudo-rewards for all $k \neq k_t$: $\hat{\phi}_{k,k_t}(t) = \sum_{\tau: k_\tau = k_t} s_{k,k_\tau}(r_\tau)/n_{k_t}(t)$
end for

Algorithm 2 C-TS Correlated TS Algorithm

Steps 1 - 10 as in C-UCB
Apply TS over arms in $\mathcal{A}_t \cup \{k^{\textbf{emp}}(t)\}$ by pulling arm $k_t = \arg\max_{k \in \mathcal{A}_t \cup \{k^{\text{emp}}(t)\}} S_{k,t}$, where $S_{k,t} \sim \mathcal{N}\left(\hat{\mu}_k(t), \frac{\beta B}{n_k(t)+1}\right)$.
Receive reward r_t, and update $n_{k_t}(t)$, $\hat{\mu}_{k_t}(t)$ and empirical pseudo-rewards $\hat{\phi}_{k,k_t}(t)$.

IV. Regret Analysis and Bounds

We now characterize the performance of the C-UCB algorithm by analyzing the expected value of the cumulative regret (3). The expected regret can be expressed as

$$\mathbb{E}\left[Reg(T)\right] = \sum_{k=1}^{K} \mathbb{E}\left[n_k(T)\right] \Delta_k, \tag{9}$$

where $\Delta_k = \mu_{k^*} - \mu_k$ is the sub-optimality gap of arm k with respect to the optimal arm k^*, and $n_k(T)$ is the number of times arm k is pulled in T slots.

For the regret analysis, we assume without loss of generality that the rewards are between 0 and 1 for all $k \in \{1, 2, \dots K\}$. Note that the C-BANDIT algorithms do not require this condition, and the regret analysis can also be generalized to any bounded rewards.

A. Competitive and Non-competitive arms with respect to Arm k

For the purpose of regret analysis in Section IV, we need to understand which arms are empirically competitive as $t \to \infty$.

We do so by defining the notions of Competitive and Non-Competitive arms.

Definition 5 (Non-Competitive and Competitive arms)**.** *An arm ℓ is said to be non-competitive if the expected reward of optimal arm k^* is larger than the expected pseudo-reward of arm ℓ with respect to the optimal arm k^*, i.e, if, $\tilde{\Delta}_{\ell,k^*} \triangleq \mu_{k^*} - \phi_{\ell,k^*} > 0$. Similarly, an arm ℓ is said to be competitive if $\tilde{\Delta}_{\ell,k^*} = \mu_{k^*} - \phi_{\ell,k^*} <= 0$. The unique best arm k^* has $\tilde{\Delta}_{k^*,k^*} = \mu_{k^*} - \phi_{k^*,k^*} = 0$ and is counted in the set of competitive arms.*[3]

We refer to $\tilde{\Delta}_{\ell,k^*}$ as the pseudo-gap of arm ℓ in the rest of the paper. These notions of competitiveness are used in the regret analysis in Section IV. The central idea behind our correlated C-BANDIT approach is that after pulling the optimal arm k^* sufficiently large number of times, the non-competitive (and thus sub-optimal) arms can be classified as empirically non-competitive with increasing confidence, and thus need not be explored. As a result, the non-competitive arms will be pulled only $\mathrm{O}(1)$ times. However, the competitive arms cannot be discerned as sub-optimal by just using the rewards observed from the optimal arm, and have to be explored $\mathrm{O}(\log T)$ times each. Thus, we are able to reduce a K-armed bandit to a C-armed bandit problem, where C is the number of competitive arms. [4] We show this by bounding the regret of C-BANDIT approach.

B. Regret Bounds

In order to bound $\mathbb{E}\left[Reg(T)\right]$ in (9), we can analyze the expected number of times sub-optimal arms are pulled, that is, $\mathbb{E}\left[n_k(T)\right]$, for all $k \neq k^*$. Theorem 1 and Theorem 2 below show that $\mathbb{E}\left[n_k(T)\right]$ scales as $O(1)$ and $O(\log T)$ for non-competitive and competitive arms respectively. Recall that a sub-optimal arm is said to be non-competitive if its pseudo-gap $\tilde{\Delta}_{k,k^*} > 0$, and competitive otherwise.

Theorem 1 (Expected Pulls of a Non-competitive Arm)**.** *The expected number of times a non-competitive arm with pseudo-gap $\tilde{\Delta}_{k,k^*}$ is pulled by C-UCB is upper bounded as*

$$\mathbb{E}\left[n_k(T)\right] \leq Kt_0 + K^3 \sum_{t=Kt_0}^{T} 2\left(\frac{t}{K}\right)^{-2} + \sum_{t=1}^{T} 3t^{-3}, \quad (10)$$

$$= \mathrm{O}(1), \quad (11)$$

where,

$$t_0 = \inf\left\{\tau \geq 2 : \Delta_{min}, \tilde{\Delta}_{k,k^*} \geq 4\sqrt{\frac{2K\log\tau}{\tau}}\right\}.$$

[3] As $t \to \infty$, only the optimal arm will remain in $\mathcal{S}_t$, and hence the definition of competitive arms only compares the expected mean of arm k^* and expected pseudo-reward of arm k with respect to arm k^*

[4] Observe that k^* and subsequently C are both unknown to the algorithm. Before the start of the algorithm, it is not known which arm is optimal/competitive/non-competitive. Algorithm works in an online manner by evaluating the noisy notions of competitiveness, i.e., empirically competitive arms, and ensures that only $C-1$ of the arms are pulled $\mathrm{O}(\log T)$ times.

Theorem 2 (Expected Pulls of a Competitive Arm)**.** *The expected number of times a competitive arm is pulled by C-UCB algorithm is upper bounded as*

$$\mathbb{E}\left[n_k(T)\right] \leq 8\frac{\log(T)}{\Delta_k^2} + \left(1 + \frac{\pi^2}{3}\right) + \sum_{t=1}^{T} 2Kte^{\left(-\frac{t\Delta_{min}^2}{2K}\right)}, \quad (12)$$

$$= \mathrm{O}(\log T) \quad \textit{where } \Delta_{min} = \min_k \Delta_k > 0. \quad (13)$$

Substituting the bounds on $\mathbb{E}\left[n_k(T)\right]$ derived in Theorem 1 and Theorem 2 into (9), we get the following upper bound on expected regret.

Corollary 1 (Upper Bound on Expected Regret)**.** *The expected cumulative regret of the C-UCB and C-TS algorithms is upper bounded as*

$$\mathbb{E}\left[Reg(T)\right] \leq \sum_{k\in\mathcal{C}\setminus\{k^*\}} \Delta_k U_k^{(c)}(T) + \sum_{k'\in\mathcal{K}\setminus\{\mathcal{C}\}} \Delta_{k'} U_{k'}^{(nc)}(T), \quad (14)$$

$$= (C-1)\cdot\mathrm{O}(\log T) + \mathrm{O}(1), \quad (15)$$

where $\mathcal{C} \subseteq \{1,\ldots,K\}$ is set of competitive arms with cardinality C, $U_k^{(c)}(T)$ is the upper bound on $\mathbb{E}\left[n_k(T)\right]$ for competitive arms given in (2), and $U_k^{(nc)}(T)$ is the upper bound for non-competitive arms given in (1).

C. Proof Sketch

We now present an outline of our regret analysis of C-UCB. A key strength of our analysis is that it can be extended very easily to any C-BANDIT algorithm. The results independent of last step in the algorithm are presented in Appendix B, while the rigorous regret upper bounds for C-UCB is presented in Appendix D. We also present a regret analysis for C-TS in a scenario where $K = 2$, and TS is employed with Beta priors in Appendix E.

There are three key components to prove the result in Theorem 1 and Theorem 2. The first two components hold independent of which bandit algorithm (UCB/TS/KL-UCB) is used for selecting the arm from the set of competitive arms, which makes our analysis easy to extend to any C-BANDIT algorithm. The third step is specific to the last step in C-BANDIT algorithm. We analyse the third component for C-UCB to provide its rigorous regret results.

i) Probability of optimal arm being identified as empirically non-competitive at round t (denoted by $\Pr(E_1(t))$) is small. In particular, we show that

$$\Pr(E_1(t)) \leq 2Kt\exp\left(-\frac{t\Delta_{\min}^2}{2K}\right).$$

This ensures that the optimal arm is identified as empirically non-competitive only $\mathrm{O}(1)$ times. We show that the number of times a competitive arm is pulled is bounded as

$$\mathbb{E}\left[n_k(T)\right] \leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(E_1^c(t), k_t = k, I_{k,t-1} > I_{k^*,t-1}). \quad (16)$$

The first term sums to a constant, while the second term is upper bounded by the number of times UCB pulls the sub-optimal arm k. Due to this the upper bound on the number of pulls of competitive arm by C-UCB / C-TS is only an additive constant more than the upper bound on the number of pulls for an arm by UCB / TS algorithms and hence we have same pre-log constants for the upper bound on the pulls of competitive arms.

ii) Probability of identifying a non-competitive arm as empirically competitive jointly with optimal arm being pulled more than $\frac{t}{K}$ times is small. Notice that the first two steps of our algorithm involve identifying the set of arms $\mathcal{S}_t$ that have been pulled at least $\frac{t}{K}$ times, and eliminating arms which are empirically non-competitive with respect to the set $\mathcal{S}_t$ for round t. We show that the joint event that arm $k^* \in \mathcal{S}_t$ and a non-competitive arm k is identified as empirically non-competitive is small. Formally,

$$\Pr\left(k_{t+1}=k, n_{k^*}(t) \geq \frac{t}{K}\right) \leq t \exp\left(-\frac{t\tilde{\Delta}_{k,k^*}}{2K}\right). \quad (17)$$

This occurs because upon obtaining a *large* number of samples of arm k^*, expected reward of arm k^* (i.e., μ_{k^*}) and expected pseudo-reward of arm k with respect to arm k^* (i.e., ϕ_{k,k^*}) can be estimated *fairly accurately*. Since the pseudo-gap of arm k is positive (i.e., $\mu_{k^*} > \phi_{k,k^*}$), the probability that arm k is identified as empirically competitive is small. An implication of (17) is that the expected number of times a non-competitive arm is identified as empirically competitive jointly with the optimal arm having at least $\frac{t}{K}$ pulls at round t is bounded above by a constant.

iii) **Probability that a sub-optimal arm is pulled more than t/K times at round t is small.** Formally, we show that for C-UCB, we have

$$\Pr\left(n_k(t) \geq \frac{t}{K}\right) \leq (2K+2)\left(\frac{t}{K}\right)^{-2} \quad \forall t > Kt_0, k \neq k^* \quad (18)$$

This component of our analysis is specific to the classical bandit algorithm used in C-BANDIT. Intuitively, a result of this kind should hold for any *good performing* classical multi-armed bandit algorithm. We reach the result of (18) in C-UCB by showing that

$$\Pr\left(k_{t+1}=k, n_k(t) > \frac{t}{2K}\right) \leq t^{-3} \quad \forall t > t_0, k \neq k^* \quad (19)$$

The probability of selecting a sub-optimal arm k after it has been pulled *significantly* many times is small as with more number of pulls, the exploration component in UCB index of arm k becomes small, and consequently it is likely to be smaller than the UCB index of optimal arm k^* (as it has larger empirical mean reward or has been pulled fewer number of times). Our analysis in Lemma 9 shows how the result in (19) can be translated to obtain (18) (this translation is again not dependent on which bandit algorithm is used in C-BANDIT).

$p_1(r)$	$\mathbf{r}$	$s_{2,1}(r)$	$s_{3,1}(r)$
0.2	**0**	0.7	2
0.2	**1**	0.8	1.2
0.6	**2**	2	1

TABLE III: Suppose Arm 1 is optimal and its unknown probability distribution is $(0.2, 0.2, 0.6)$, then $\mu_1 = 1.4$, while $\phi_{2,1} = 1.5$ and $\phi_{3,1} = 1.2$. Due to this Arm 2 is Competitive while Arm 3 is non-competitive

We show that the expected number of pulls of a non-competitive arm k can be bounded as

$$\mathbb{E}\left[n_k(T)\right] \leq \sum_{t=1}^{T}\left(\Pr\left(k_{t+1}=k, k^* = \arg\max_k n_k(t)\right) + \Pr\left(k^* \neq \arg\max_k n_k(t)\right)\right) \quad (20)$$

The first term in (20) is $\mathrm{O}(1)$ due to (17) and the second term is $\mathrm{O}(1)$ due to (18). Refer to Appendix D for rigorous regret analysis of C-UCB.

D. Discussion on Regret Bounds

Competitive Arms. Recall than an arm is said to be competitive if μ_{k^*} (i.e., expected reward from arm k^*) $> \mathbb{E}\left[\phi_{k,k^*}\right] = \mathbb{E}\left[\tilde{\mathbb{E}}[R_{k'}|R_k]\right]$. Since the distribution of reward of each arm is unknown, initially the Algorithm does not know which arm is *competitive* and which arm is *non-competitive*.

Reduction in effective number of arms. Interestingly, our result from Theorem 1 shows that the C-UCB algorithm, that operates in a sequential fashion, makes sure that *non-competitive* arms are pulled only $\mathrm{O}(1)$ times. Due to this, only the competitive arms are pulled $\mathrm{O}(\log T)$ times. Moreover, the pre-log terms in the upper bound of UCB and C-UCB for these arms is the same. In this sense, our C-BANDIT approach reduces a K-armed bandit problem to a C-armed bandit problem. Effectively only $C-1 \leq K-1$ arms are pulled $\mathrm{O}(\log T)$ times, while other arms are stopped being pulled after a finite time.

Depending on the joint probability distribution, different arms can be optimal, competitive or non-competitive. Table III shows a case where arm 1 is optimal and the reward distribution of arm 1 is $(0.2, 0.2, 0.6)$, which leads to $\mu_1 = 1.4 > \phi_{3,1} = 1.2$ and $\mu_1 = 1.4 < \phi_{2,1} = 1.5$. Due to this Arm 2 is competitive while Arm 3 is non-competitive.

Achieving Bounded Regret. If the set of competitive arms $\mathcal{C}$ is a singleton set containing only the optimal arm (i.e., the number of competitive arms $C = 1$), then our algorithm will lead to (see (15)) an expected regret of $\mathrm{O}(1)$, instead of the typical $\mathrm{O}(\log T)$ regret scaling in classic multi-armed bandits. One such scenarion in which this can happen is if pseudo-rewards s_{k,k^*} of all arms with respect to optimal arm k^* match the conditional expectation of arm k. Formally, if $s_{k,k^*} = \mathbb{E}\left[R_k|R_{k^*}\right] \forall k$, then $\mathbb{E}\left[s_{k,k^*}\right] = \mathbb{E}\left[R_k\right] = \mu_k < \mu_{k^*}$. Due to this, all sub-optimal arms are non-competitive and our algorithms achieve only $\mathrm{O}(1)$ regret. We now evaluate a lower bound result for a special case of our model, where rewards

are correlated through a latent random variable X as described in Section II-C.

We present a lower bound on the expected regret for the model described in Section II-C. Intuitively, if an arm ℓ is *competitive*, it can not be deemed sub-optimal by only pulling the optimal arm k^* infinitely many times. This indicates that exploration is necessary for competitive sub-optimal arms. The proof of this bound closely follows that of the 2-armed classical bandit problem [1]; i.e., we construct a new bandit instance under which a previously sub-optimal arm becomes optimal without affecting reward distribution of any other arm.

Theorem 3 (Lower Bound for Correlated MAB with latent random source)**.** *For any algorithm that achieves a sub-polynomial regret, the expected cumulative regret for the model described in Section II-C is lower bounded as*

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[Reg(T)\right]}{\log(T)} \geq \begin{cases} \max_{k\in\mathcal{C}} \frac{\Delta_k}{D(f_{R_k}||f_{\tilde{R}_k})} & \text{if } C > 1 \\ 0 & \text{if } C = 1. \end{cases} \tag{21}$$

Here f_{R_k} is the reward distribution of arm k, which is linked with f_X since $R_k = Y_k(X)$. The term $f_{\tilde{R}_k}$ represents the reward distribution of arm k in the new bandit instance where arm k becomes optimal and distribution $f_{R_{k^*}}$ is unaffected. The divergence term represents "the amount of distortion needed in reward distribution of arm k to make it better than arm k^*", and hence captures the problem difficulty in the lower bound expression.

Bounded regret whenever possible for the special case of Section II-C. From Corollary 1, we see that whenever $C > 1$, our proposed algorithm achieves $\mathrm{O}(\log T)$ regret matching the lower bound given in Theorem 3 order-wise. Also, when $C = 1$, our algorithm achieves $\mathrm{O}(1)$ regret. Thus, our algorithm achieves bounded regret whenever possible, i.e., when $C = 1$ for the model described in Section II-C. In the general problem setting, a lower bound $\Omega(\log T)$ exists whenever it is possible to change the joint distribution of rewards such that the marginal distribution of optimal arm k^* is unaffected and pseudo-rewards $s_{\ell,k}(r)$ still remain an upper bound on $\mathbb{E}\left[R_\ell | R_k = r\right]$ under the new joint probability distribution. In general, this can happen even if $C = 1$, we discuss one such scenario in the Appendix F and explain the challenges that need to come from the algorithmic side to meet the lower bound.

V. Simulations

We now present the empirical performance of proposed algorithms. For all the results presented in this section, we compare the performance of all algorithms on the same reward realizations and plot the cumulative regret averaged over 100 independent trials. The shaded area represents error bars with one standard deviation. We set $\beta = 1$ for all TS and C-TS plots.

A. Simulations with known pseudo-rewards

Consider the example shown in Table I, with the top row showing the pseudo-rewards, which are known to the player,

r	$s_{2,1}(r)$
0	0.7
1	0.4

r	$s_{1,2}(r)$
0	0.8
1	0.5

(a)	$R_1 = 0$	$R_1 = 1$
$R_2 = 0$	0.2	0.4
$R_2 = 1$	0.2	0.2

(b)	$R_1 = 0$	$R_1 = 1$
$R_2 = 0$	0.2	0.3
$R_2 = 1$	0.4	0.1

TABLE IV: The top row shows the pseudo-rewards of arms 1 and 2, i.e., upper bounds on the conditional expected rewards (which are known to the player). The bottom row depicts two possible joint probability distribution (unknown to the player). Under distribution (a), Arm 1 is optimal whereas Arm 2 is optimal under distribution (b).

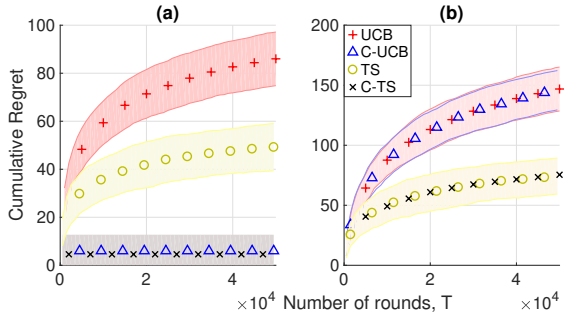

Fig. 5: Cumulative regret for UCB, C-UCB, TS and C-TS corresponding to the problem shown in Table IV. For the setting (a) in Table IV, Arm 1 is optimal and Arm 2 is non-competitive, in setting (b) of Table IV Arm 2 is optimal while Arm 1 is competitive.

and the bottom row showing two possible joint probability distributions (a) and (b), which are unknown to the player. We show the simulation result of our proposed algorithms C-UCB, C-TS against UCB, TS in Figure 5 for the setting considered in Table I.

Case (a): Bounded regret. For the probability distribution (a), notice that Arm 1 is optimal with $\mu_1 = 0.6, \mu_2 = 0.4$. Moreover, $\phi_{2,1} = 0.4\times0.7+0.6\times0.4 = 0.52$. Since $\phi_{2,1} < \mu_1$, Arm 2 is non-competitive. Hence, in Figure 5(a), we see that our proposed C-UCB and C-TS Algorithms achieve bounded regret, whereas UCB, TS show logarithmic regret.

Case (b): All competitive arms. For the probability distribution (b), Arm 2 is optimal with $\mu_2 = 0.5$ and $\mu_1 = 0.4$. The expected pseudo-reward of arm 1 w.r.t to arm 2 in this case is $\phi_{1,2} = 0.8 \times 0.5 + 0.5 \times 0.5 = 0.65$. Since $\phi_{1,2} > \mu_2$, the sub-optimal arm (i.e., Arm 1) is competitive and hence C-UCB and C-TS also end up exploring Arm 1. Due to this we see that C-UCB, C-TS achieve a regret similar to UCB, TS in Figure 5(b). C-TS has empirically smaller regret than C-UCB as Thompson Sampling performs better empirically than the UCB algorithm. The design of our C-Bandit approach allows the use of any other bandit algorithm in the last step, e.g., KL-UCB.

B. Simulations for the latent random source model in Section II-C

We now show the performance of C-UCB and C-TS against UCB, TS for the model considered in Section II-C, where rewards corresponding to different arms are correlated through

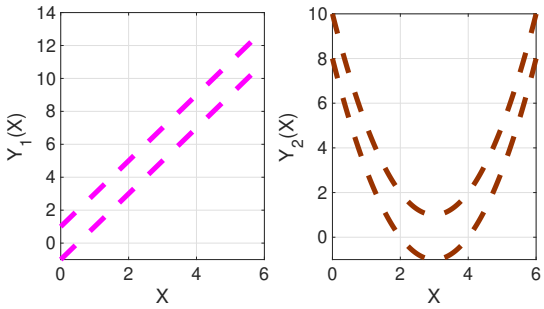

Fig. 6: Rewards corresponding to the two arms are correlated through a random variable X lying in $(0, 6)$. The lines represent lower and upper bounds on reward of Arms 1 ,$Y_1(X)$, and 2, $Y_2(X)$, given the realization of random variable X.

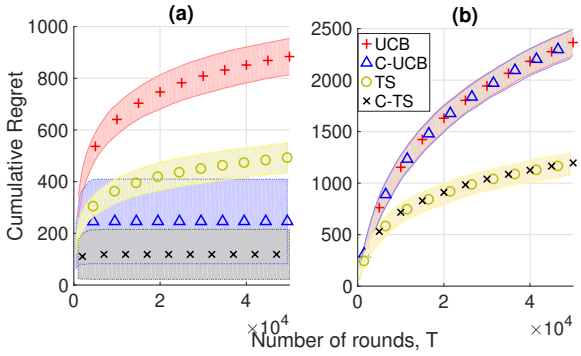

Fig. 7: Simulation results for the example shown in Figure 6. In (a), $X \sim \text{Beta}(1, 1)$ and in (b) $X \sim \text{Beta}(1.5, 5)$. In case (a), Arm 1 is optimal while Arm 2 is non-competitive (C = 1), due to which we see that C-UCB and C-TS obtain bounded regret. Arm 2 is optimal for the distribution in (b) and Arm 1 is competitive, due to which $C = 2$ and we see that C-UCB and C-TS attain a performance similar to UCB and TS.

a latent random variable X. We consider a setting where reward obtained from Arm 1, given a realization x of X, is bounded between $2x - 1$ and $2x + 1$, i.e., $2X - 1 \leq Y_1(X) \leq 2X + 1$. Similarly, conditional reward of Arm 2 is, $(3 - X)^2 - 1 \leq Y_2(X) \leq (3 - X)^2 + 1$. Figure 6 demonstrates these upper and lower bounds on $Y_k(X)$. We run C-UCB, C-TS, TS and UCB for this setting for two different distributions of X. For the simulations, we set the conditional reward of both the arms to be distributed uniformly between the upper and lower bounds, however this information is not known to the Algorithms.

Case (a): $X \sim$ **Beta**$(1, 1)$. When X is distributed as $X \sim \text{Beta}(1, 1)$, Arm 1 is optimal while Arm 2 is non-competitive. Due to this, we observe that C-UCB and C-TS achieve bounded regret in Figure 7(a).

Case (b): $X \sim$ **Beta**$(1.5, 5)$. In the scenario where X has the distribution $\text{Beta}(1.5, 5)$, Arm 2 is optimal while Arm 1 is competitive. Due to this, C-UCB and C-TS do not stop exploring Arm 1 in finite time and we see the cumulative regret similar to UCB, TS in Figure 7(b).

Our next simulation result considers a setting where the known upper and lower bounds on $Y_k(X)$ match and the reward Y_k corresponding to a realization of X is deterministic,

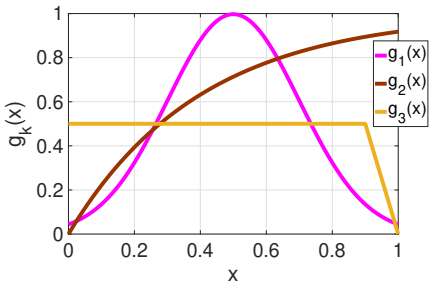

Fig. 8: Reward Functions used for the simulation results presented in Figure 9. The reward $g_k(X)$ is a function of a latent random variable X. For instance, when $X = 0.5$, reward from Arms 1,2 and 3 are $g_1(X) = 1$, $g_2(X) = 0.7135$ and $g_3(X) = 0.5$.

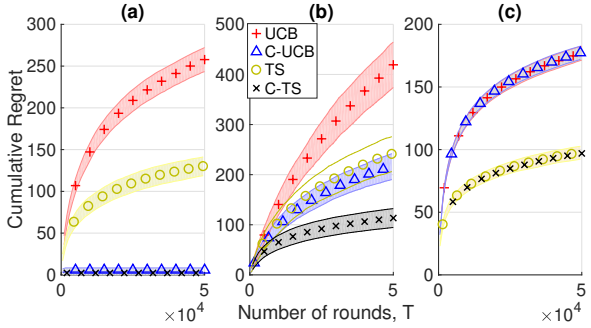

Fig. 9: *The cumulative regret of C-UCB and C-TS depend on C, the number of* competitive *arms. The value of C depends on the* unknown *joint probability distribution of rewards and is not known beforehand. We consider a setup where $C = 1$ in (a), $C = 2$ in (b) and $C = 3$ in (c). Our proposed algorithm pull only the $C - 1$ competitive sub-optimal arms $\text{O}(\log T)$ times, as opposed to UCB, TS that pull all $K - 1$ sub-optimal arms $\text{O}(\log T)$ times. Due to this, we see that our proposed algorithms achieve bounded regret when $C = 1$. When $C = 3$, our proposed algorithms perform as well as the UCB, TS algorithms.*

i.e., $Y_k(X) = g_k(X)$. We show our simulation results for the reward functions described in Figure 8 with three different distributions of X. Corresponding to $X \sim \text{Beta}(4, 4)$, Arm 1 is optimal and Arms 2,3 are non-competitive leading to bounded regret for C-UCB, C-TS in Figure 9(a). In setting (b), we consider $X \sim \text{Beta}(2, 5)$ in which Arm 1 is optimal, Arm 2 is competitive and Arm 3 is non-competitive. Due to this, our proposed C-UCB and C-TS Algorithms stop pulling Arm 3 after some time and hence achieve significantly reduced regret relative to UCB in Figure 9(b). For third scenario (c), we set $X \sim \text{Beta}(1, 5)$, which makes Arm 3 optimal while Arms 1 and 2 are competitive. Hence, our algorithms explore both the sub-optimal arms and have a regret comparable to that of UCB, TS in Figure 9(c).

VI. Experiments

We now show the performance of our proposed algorithms in real-world settings. Through the use of MovieLens and Goodreads datasets, we demonstrate how the correlated

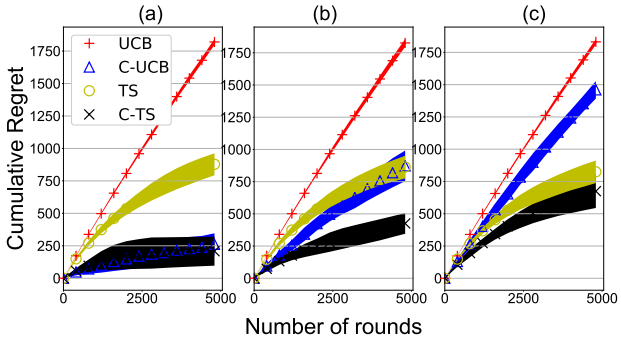

Fig. 10: Cumulative regret for UCB, C-UCB, TS and C-TS for the application of recommending the best genre in the Movielens dataset, where p fraction of the pseudo-entries are replaced with maximum reward $i.e., 5$. In $(a), p = 0.25$, for $(b), p = 0.50$ and $p = 0.7$ in (c). The value of C is $4, 11$ and 13 in $(a), (b)$ and (c) respectively. As C is smaller than K (i.e., 18) in each case, we see that C-UCB and C-TS outperform UCB and TS significantly.

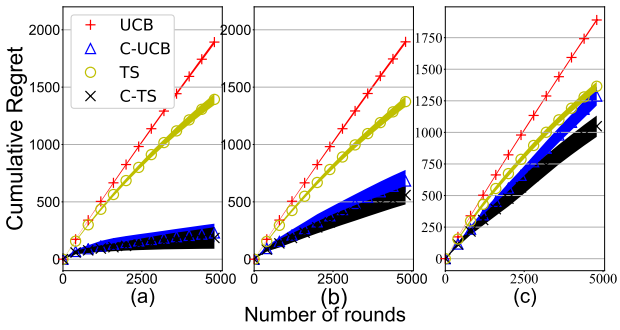

Fig. 11: Cumulative regret of UCB, C-UCB, TS and C-TS for providing the best movie recommendations in the Movielens dataset. Each pseudo-reward entry is added by 0.1 in (a), 0.4 in (b) and 0.6 in (c). The value of C is $6, 24$ and 39 in $(a), (b)$ and (c) respectively. As C is smaller than K (i.e., 50) in each case, we see the superior performance of C-UCB, C-TS over UCB and TS.

MAB framework can be used in practical settings for recommendation system applications. In such systems, it is possible to use the prior available data (from a certain population) to learn the correlation structure in the form of pseudo-rewards. When trying to design a campaign to maximize user engagement in a new unknown demographic, the learned correlation information in the form of pseudo-rewards can help significantly reduce the regret as we show from our results described below.

A. *Experiments on the* MovieLens *dataset*

The MovieLens dataset [30] contains a total of 1M ratings for a total of 3883 Movies rated by 6040 Users. Each movie is rated on a scale of 1-5 by the users. Moreover, each movie is associated with one (and in some cases, multiple) genres. For our experiments, of the possibly several genres associated with each movie, one is picked uniformly at random. To perform our experiments, we split the data into two parts, with the first half containing ratings of the users who provided the most number of ratings. This half is used to learn the pseudo-reward entries, the other half is the test set which is used to evaluate the performance of the proposed algorithms. Doing such a split ensures that the rating distribution is different in the training and test data.

Recommending the Best Genre. In our first experiment, the goal is to provide the best genre recommendations to a population with unknown demographic. We use the training dataset to learn the pseudo-reward entries. The pseudo-reward entry $s_{\ell,k}(r)$ is evaluated by taking the empirical average of the ratings of genre ℓ that are rated by the users who rated genre k as r. To capture the fact that it might not be possible in practice to fill all pseudo-reward entries, we randomly remove p-fraction of the pseudo-reward entries. The removed pseudo-reward entries are replaced by the maximum possible rating, i.e., 5 (as that gives a natural upper bound on the conditional mean reward). Using these pseudo-rewards, we evaluate our proposed algorithms on the test data. Upon recommending a particular genre (arm), the rating (reward) is obtained by sampling one of the ratings for the chosen arm in the test data. Our experimental results for this setting are shown in Figure 10, with $p = 0.25, 0.50$ and 0.70 (i.e., fraction of pseudo-reward entries that are removed). We see that the proposed C-UCB and C-TS algorithms significantly outperform UCB and TS in all three settings. For each of the three cases we also evaluate the value of C (which is unknown to the algorithm), by always pulling the optimal arm and finding the size of empirically competitive set at $T = 5000$. The value of C turned out to be $4, 11$ and 13 for $p = 0.25, 0.50$ and 0.70. As $C < 18$ in each case, some of the 18 arms are stopped being pulled after some time and due to this, C-UCB and C-TS significantly outperform UCB and TS respectively. This shows that even when only a subset of the correlations are known, it is possible to exploit them to improve the performance of classical bandit algorithms.

Recommending the Best Movie. We now consider the goal of providing the best movie recommendations to the population. To do so, we consider the 50 most rated movies in the dataset. containing 109,804 user-ratings given by 6,025 users. In the testing phase, the goal is to recommend one of these 50 movies to each user. As was the case in previous experiment, we learn the pseudo-reward entries from the training data. Instead of using the learned pseudo-reward directly, we add a *safety buffer* to each of the pseudo-reward entry; i.e., we set the pseudo-reward as the empirical conditional mean *plus* the SAFETY BUFFER. Adding a buffer will be needed in practice, as the conditional expectations learned from the training data are likely to have some noise and adding a safety buffer allows us to make sure that pseudo-rewards constitute an upper bound on the conditional expectations. Our experimental result in Figure 11 shows the performance of C-UCB and C-TS relative to UCB for this setting with safety buffer set to 0.1 in Figure 11(a), 0.4 in Figure 11(b) and 0.6 in Figure 11(c). In all three cases, even after addition of safety buffers, our proposed C-UCB and C-TS algorithms outperform the UCB algorithm.

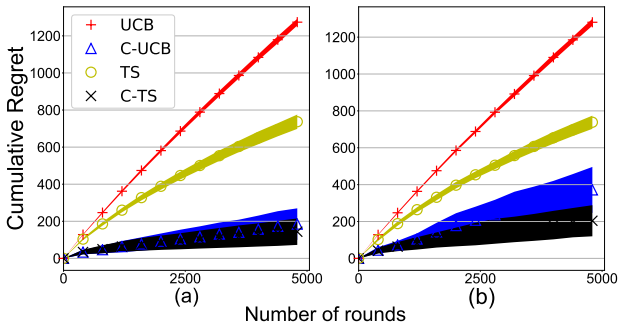

Fig. 12: Cumulative regret of UCB, C-UCB, TS and C-TS for providing best poetry book recommendation in the Goodreads dataset. Every pseudo-reward entry is added by q and p fraction of the pseudo-reward entries are removed, with (a) $p = 0.1, q = 0.1$ and (b) $p = 0.3, q = 0.1$. The value of C is 8 and 11 in (a) and (b) respectively. As C is much smaller than K (i.e., 25) in each case, we see that C-UCB and C-TS outperform UCB and TS significantly.

B. Experiments on the GOODREADS dataset

The GOODREADS dataset [31] contains the ratings for 1,561,465 books by a total of 808,749 users. Each rating is on a scale of 1-5. For our experiments, we only consider the poetry section and focus on the goal of providing best poetry recommendations to the whole population whose demographics is unknown. The poetry dataset has 36,182 different poems rated by 267,821 different users. We do the pre-processing of goodreads dataset in the same manner as that of the MovieLens dataset, by splitting the dataset into two halves, train and test. The train dataset contains the ratings of the users with most number of recommendations.

Recommending the best poetry book. We consider the 25 most rated books in the dataset and use these as the set of arms to recommend in the testing phase. These 25 poems have 349,523 user-ratings given by 171,433 users. As with the MOVIELENS dataset, the pseudo-reward entries are learned on the training data. In practical situations it might not be possible to obtain all pseudo-reward entries. Therefore, we randomly select p fraction of the pseudo-reward entries and replace them with maximum possible reward (i.e. 5). Among the remaining pseudo-reward entries we add a safety buffer of q to each entry. Our result in Figure 12 shows the performance of C-UCB and C-TS relative to UCB and TS in two scenarios. In scenario (a), 10% of the pseudo-reward entries are replaced by 5 and remaining are padded with a safety buffer of 0.1. For case (b), 30% entries are replaced by 5 and safety buffer is 0.1. Under both cases, our proposed C-UCB and C-TS algorithms are able to outperform UCB and TS significantly.

C. Pseudo-rewards learned on a smaller dataset

In our previous set of experiments, half of the dataset was used to learn the pseudo-reward entries. We did additional experiments in a setup where only 10% of the data was used for learning the pseudo-reward entries and tested our algorithms on the remaining dataset. On doing so, we observed that C-UCB and C-TS were still able to outperform UCB and TS in most of our experimental setups. One setting in which the performance of C-UCB was similar to that of UCB is in a scenario where each pseudo-reward entry was padded by 0.6. As the padding was large, the C-UCB algorithm was not able to identify many arms as non-competitive, leading to a performance that is similar to that of UCB. In all other scenarios, we noted that C-UCB and C-TS significantly outperformed UCB and TS, suggesting that even when smaller dataset is used for learning pseudo-rewards, the C-UCB and C-TS can be quite effective. The results are presented in Figure 13, Figure 14 and Figure 15.

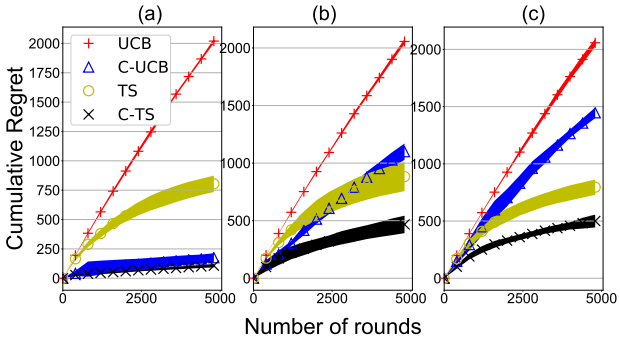

Fig. 13: Cumulative regret for UCB, C-UCB, TS and C-TS for the application of recommending the best genre in the Movielens dataset, where p fraction of the pseudo-entries are replaced with maximum reward $i.e., 5$. In $(a), p = 0.25$, for $(b), p = 0.50$ and $p = 0.7$ in (c). We used 10% of the dataset to learn the pseudo-reward entry and the algorithms are tested on the remaining dataset. The value of C is $5, 11$ and 15 in $(a), (b)$ and (c) respectively. As C is smaller than K (i.e., 18) in each case, we see that C-UCB and C-TS outperform UCB and TS significantly. Note that the value of C is larger in the case where only 10% data is used for learning the pseudo-reward.

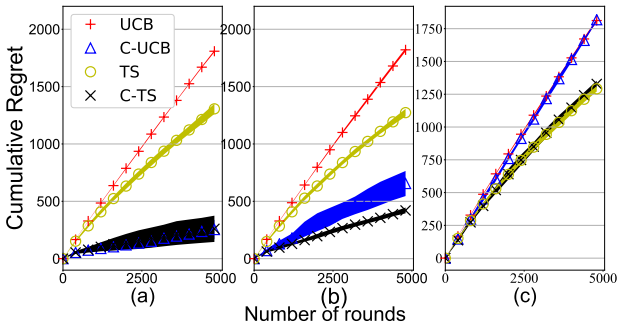

Fig. 14: Cumulative regret of UCB, C-UCB, TS and C-TS for providing the best movie recommendations in the Movielens dataset. In this experiment 10% of the dataset is used for learning the pseudo-reward entry and the algorithms are tested on the remaining dataset. Each pseudo-reward entry is added by 0.1 in (a), 0.4 in (b) and 0.6 in (c). The value of C is $14, 29$ and 42 in $(a), (b)$ and (c) respectively. Note that the value of C is larger in the case where only 10% data is used for learning the pseudo-reward. As C is still smaller than K (i.e., 50) in each case, we see the superior performance of C-UCB, C-TS over UCB and TS.

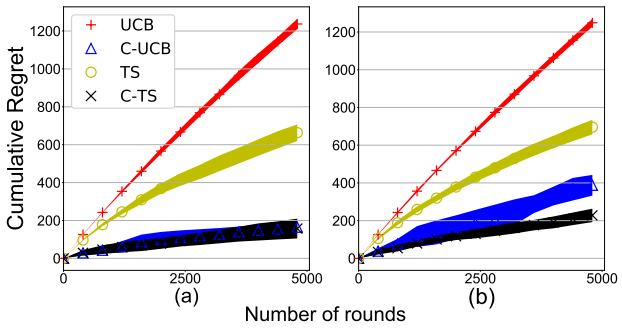

Fig. 15: Cumulative regret of UCB, C-UCB, TS and C-TS for providing best poetry book recommendation in the Goodreads dataset. We used 10% of the dataset to learn the pseudo-reward entry and the algorithms are tested on the remaining dataset. Every pseudo-reward entry is added by q and p fraction of the pseudo-reward entries are removed, with (a) $p = 0.1, q = 0.1$ and (b) $p = 0.3, q = 0.1$. The value of C is 7 and 12 in (a) and (b) respectively. As C is much smaller than K (i.e., 25) in each case, we see that C-UCB and C-TS outperform UCB and TS significantly.

VII. Conclusion

This work presents a new correlated Multi-Armed bandit problem in which rewards obtained from different arms are correlated. We capture this correlation through the knowledge of *pseudo-rewards*. These pseudo-rewards, which represent upper bound on conditional mean rewards, could be known in practice from either domain knowledge or learned from prior data. Using the knowledge of these pseudo-rewards, we the propose *C-Bandit* algorithm which fundamentally generalizes any classical bandit algorithm to the correlated multi-armed bandit setting. A key strength of our paper is that it allows pseudo-rewards to be loose (in case there is not much prior information) and even then our *C-Bandit* algorithms adapt and provide performance at least as good as that of classical bandit algorithms.

We provide a unified method to analyze the regret of C-Bandit algorithms. In particular, the analysis shows that C-UCB ends up pulling *non-competitive* arms only $\mathrm{O}(1)$ times; i.e., they stop pulling certain arms after a finite time t. Due to this, C-UCB pulls only $C - 1 \leq K - 1$ of the $K - 1$ sub-optimal arms $\mathrm{O}(\log T)$ times, as opposed to UCB that pulls *all* $K - 1$ sub-optimal arms $\mathrm{O}(\log T)$ times. In this sense, our C-Bandit algorithms reduce a K-armed bandit to a C-armed bandit problem. We present several cases where $C = 1$ for which C-UCB achieves bounded regret. For the special case when rewards are correlated through a latent random variable X, we provide a lower bound showing that bounded regret is possible only when $C = 1$; if $C > 1$, then $\mathrm{O}(\log T)$ regret is not possible to avoid. Thus, our C-UCB algorithm achieves bounded regret whenever possible. Simulation results validate the theoretical findings and we perform experiments on MovieLens and Goodreads datasets to demonstrate the applicability of our framework in the context of recommendation systems. The experiments on real-world datasets show that our C-UCB and C-TS algorithms significantly outperform the UCB and TS algorithms.

There are several interesting open problems and extensions of this work, some of which we describe below.

Extension to light tailed and heavy tailed rewards In this work, we assume that the rewards have a bounded support. The algorithm and analysis can be extended to settings with sub-gaussian rewards as well. In particular, in step 3 of the algorithm, one would play UCB/TS for sub-gaussian rewards. For instance, the UCB index in the scenario of sub-gaussian rewards can be redefined as $\hat{\mu}_k + \sqrt{\frac{2\sigma^2 \log t}{n_k(t)}}$, where σ is the sub-Gaussianity parameter of the reward distribution. Similar regret bounds will hold in this setting as well because the Hoeffding's inequality used in our regret analysis is valid for sub-Gaussian rewards as well. For heavy-tailed rewards, the Hoeffding's inequality is not valid. Due to which, one would need to construct confidence bounds for UCB in a different manner as done in [32]. On doing so, the C-Bandit algorithm can be employed in heavy-tailed reward settings. However, the regret analysis may not extend directly as one would need to use modified concentration inequalities to obtain bounds on mean reward of arm k as done in Lemma 1 of [32].

Designing better algorithms. While our proposed algorithms are order-optimal for the model in Section 2.3, they do not match the pre-log constants in the lower bound of the regret. It may be possible to design algorithms that have smaller pre-log constants in their regret upper bound. Further discussion along these lines is presented in Appendix F. A key advantage of our approach is that our algorithms are easy to implement and they incorporate the classical bandit algorithms nicely for the problem of correlated multi-armed bandits.

Best-Arm Identification. We plan to study the problem of best-arm identification in the correlated multi-armed bandit setting, i.e., to identify the best arm with a confidence $1 - \delta$ in as few samples as possible. Since rewards are correlated with each other, we believe the sample complexity can be significantly improved relative to state of the art algorithms, such as LIL-UCB [33], [34], which are designed for classical multi-armed bandits. Another open direction is to improve the C-Bandit algorithm to make sure that it achieves bounded regret whenever possible in the general framework studied in this paper.

VIII. Acknowledgments

This work was partially supported by the NSF through grants CCF-1840860 and CCF-2007834, the Siebel Energy Institute, the Carnegie Bosch Institute, the Manufacturing Futures Initiative, and the CyLab IoT Initiative. In addition, Samarth Gupta was supported by the CyLab Presidential Fellowship and the David H. Barakat and LaVerne Owen-Barakat CIT Dean's Fellowship.

References

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory*, 2012, pp. 39–1.

[3] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th annual Conference On Learning Theory*, 2011, pp. 359–376.

[4] S. S. Villar, J. Bowden, and J. 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*, vol. 30, no. 2, p. 199, 2015.

[5] D. Agarwal, B.-C. Chen, and P. Elango, "Explore/exploit schemes for web content optimization," in *2009 Ninth IEEE International Conference on Data Mining*. IEEE, 2009, pp. 1–10.

[6] C. Tekin and E. Turğay, "Multi-objective contextual multi-armed bandit with a dominant objective," *IEEE Transactions on Signal Processing*, vol. 66, no. 14, pp. 3799–3813, 2018.

[7] J. Nino-Mora, *Stochastic scheduling. In Encyclopedia of Optimization*, 2nd ed. New York: Springer, 2009, pp. 3818–3824.

[8] S. Krishnasamy, R. Sen, R. Johari, and S. Shakkottai, "Regret of queueing bandits," *CoRR*, vol. abs/1604.06377, 2016. [Online]. Available: http://arxiv.org/abs/1604.06377

[9] G. Joshi, "Efficient Redundancy Techniques to Reduce Delay in Cloud Systems," Ph.D. dissertation, Massachusetts Institute of Technology, Jun. 2016.

[10] J. White, *Bandit algorithms for website optimization*. " O'Reilly Media, Inc.", 2012.

[11] L. Zhou, "A survey on contextual multi-armed bandits," *CoRR*, vol. abs/1508.03326, 2015. [Online]. Available: http://arxiv.org/abs/1508.03326

[12] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2014, pp. 1638–1646.

[13] R. Combes, S. Magureanu, and A. Proutiere, "Minimal exploration in structured stochastic bandits," in *Advances in Neural Information Processing Systems*, 2017, pp. 1763–1771.

[14] T. Lattimore and R. Munos, "Bounded regret for finite-armed structured bandits," in *Advances in Neural Information Processing Systems*, 2014, pp. 550–558.

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

[16] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proceedings on the Conference on Learning Theory (COLT)*, 2008.

[17] S. Gupta, G. Joshi, and O. Yağan, "Correlated multi-armed bandits with a latent random source," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3572–3576.

[18] S. Gupta, S. Chaudhari, S. Mukherjee, G. Joshi, and O. Yağan, "A unified approach to translate classical bandit algorithms to the structured bandit setting," *IEEE Journal on Selected Areas in Information Theory*, 2020.

[19] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," in *Advances in Neural Information Processing Systems*, 2010, pp. 586–594.

[20] S. Magureanu, R. Combes, and A. Proutiere, "Lipschitz bandits: Regret lower bound and optimal algorithms," in *Conference on Learning Theory*. PMLR, 2014, pp. 975–999.

[21] Z. Wang, R. Zhou, and C. Shen, "Regional multi-armed bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 510–518.

[22] S. Pandey, D. Chakrabarti, and D. Agarwal, "Multi-armed bandit problems with dependent arms," in *Proceedings of the International Conference on Machine Learning*, 2007, pp. 721–728.

[23] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *International Conference on Machine Learning*, 2014, pp. 521–529.

[24] C. Trinh, E. Kaufmann, C. Vernade, and R. Combes, "Solving bernoulli rank-one bandits with unimodal thompson sampling," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 862–889.

[25] M. Qureshi and C. Tekin, "Fast learning for dynamic resource allocation in ai-enabled radio networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 95–110, 2019.

[26] M. A. Qureshi and C. Tekin, "Online bayesian learning for rate selection in millimeter wave cognitive radio networks," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1449–1458.

[27] H. Gupta, A. Eryilmaz, and R. Srikant, "Link rate selection using constrained thompson sampling," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 739–747.

[28] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[29] S. Agrawal and N. Goyal, "Further optimal regret bounds for thompson sampling," in *Artificial Intelligence and Statistics*, 2013, pp. 99–107.

[30] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, 4, Article 19, 2015.

[31] M. Wan and J. McAuley, "Item recommendation on monotonic behavior chains," in *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 2018, pp. 86–94.

[32] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7711–7717, 2013.

[33] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *Proceedings on the Annual Conference on Information Sciences and Systems (CISS)*, March 2014, pp. 1–6.

[34] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil'ucb: An optimal exploration algorithm for multi-armed bandits," in *Conference on Learning Theory*, 2014, pp. 423–439.

[35] T. Lattimore, "Instance dependent lower bounds," http://banditalgs.com/2016/09/30/instance-dependent-lower-bounds/.

[36] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[37] J. Bretagnolle and C. Huber, "Estimation des densitiés: risque minimax. z. für wahrscheinlichkeitstheorie und verw," *Geb.*, vol. 47, pp. 199–137, 1979.

[38] B. Van Parys and N. Golrezaei, "Optimal learning for structured bandits," 2020.

[39] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws incontrolled markov chains," *SIAM journal on control and optimization*, vol. 35, no. 3, pp. 715–743, 1997.

APPENDIX

A. Standard Results from Previous Works

Fact 1 (Hoeffding's inequality)**.** *Let $Z_1, Z_2 \dots Z_n$ be i.i.d random variables bounded between $[a,b] : a \leq Z_i \leq b$, then for any $\delta > 0$, we have*

$$\Pr\left(\left|\frac{\sum_{i=1}^{n} Z_i}{n} - \mathbb{E}\left[Z_i\right]\right| \geq \delta\right) \leq \exp\left(\frac{-2n\delta^2}{(b-a)^2}\right).$$

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 μ_k denotes the mean reward of arm k, then we have*

$$\Pr\left(\hat{\mu}_{k,n_k(t)} - \mu_k \geq \epsilon, \tau_2 \geq n_k(t) \geq \tau_1\right) \leq \sum_{s=\tau_1}^{\tau_2} \exp\left(-2s\epsilon^2\right).$$

Proof. Let $Z_1, Z_2, \dots Z_t$ 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,

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

$$\Pr\left(\left(\frac{\sum_{i=1}^{n_k(t)} Z_i}{n_k(t)} - \mu_k \geq \epsilon\right), \tau_2 \geq n_k(t) \geq \tau_1\right) \tag{22}$$

$$\leq \Pr\left(\left(\bigcup_{m=\tau_1}^{\tau_2} \frac{\sum_{i=1}^{m} Z_i}{m} - \mu_k \geq \epsilon\right), \tau_2 \geq n_k(t) \geq \tau_1\right) \tag{23}$$

$$\leq \Pr\left(\bigcup_{m=\tau_1}^{\tau_2} \frac{\sum_{i=1}^{m} Z_i}{m} - \mu_k \geq \epsilon\right) \tag{24}$$

$$\leq \sum_{s=\tau_1}^{\tau_2} \exp\left(-2s\epsilon^2\right). \tag{25}$$

□

Lemma 2 (From Proof of Theorem 1 in [28])**.** *Let $I_k(t)$ denote the UCB index of arm k at round t, and $\mu_k = \mathbb{E}\left[g_k(X)\right]$ denote the mean reward of that arm. Then, we have*

$$\Pr(\mu_k > I_k(t)) \leq t^{-3}.$$

Observe that this bound does not depend on the number $n_k(t)$ of times arm k is pulled. UCB index is defined in equation (6) of the main paper.

Proof. This proof follows directly from [28]. We present the proof here for completeness as we use this frequently in the paper.

$$\Pr(\mu_k > I_k(t)) = \Pr\left(\mu_k > \hat{\mu}_{k,n_k(t)} + \sqrt{\frac{2\log t}{n_k(t)}}\right) \tag{26}$$

$$\leq \sum_{m=1}^{t} \Pr\left(\mu_k > \hat{\mu}_{k,m} + \sqrt{\frac{2\log t}{m}}\right) \tag{27}$$

$$= \sum_{m=1}^{t} \Pr\left(\hat{\mu}_{k,m} - \mu_k < -\sqrt{\frac{2\log t}{m}}\right) \tag{28}$$

$$\leq \sum_{m=1}^{t} \exp\left(-2m\frac{2\log t}{m}\right) \tag{29}$$

$$= \sum_{m=1}^{t} t^{-4} \tag{30}$$

$$= t^{-3}. \tag{31}$$

where (27) follows from the union bound and is a standard trick (Lemma 1) to deal with random variable $n_k(t)$. We use this trick repeatedly in the proofs. We have (29) from the Hoeffding's inequality. □

Lemma 3. *Let $\mathbb{E}\left[\mathbb{1}_{I_k > I_{k^*}}\right]$ be the expected number of times $I_k(t) > I_{k^*}(t)$ in T rounds. Then, we have*

$$\mathbb{E}\left[\mathbb{1}_{I_k > I_{k^*}}\right] = \sum_{t=1}^{T} Pr(I_k > I_{k^*}) \leq \frac{8\log(T)}{\Delta_k^2} + \left(1 + \frac{\pi^2}{3}\right).$$

The proof follows the analysis in Theorem 1 of [28]. The analysis of $Pr(I_k > I_{k^*})$ is done by evaluating the joint probability $\Pr\left(I_k(t) > I_{k^*}(t), n_k(t) \geq \frac{8\log T}{\Delta_k^2}\right)$. Authors in [28] show that the probability of pulling arm k jointly with the event that it has had at-least $\frac{8\log T}{\Delta_k^2}$ pulls decays down with t, i.e., $\Pr\left(I_k(t) > I_{k^*}(t), n_k(t) \geq \frac{8\log T}{\Delta_k^2}\right) \leq t^{-2}$.

Lemma 4 (Theorem 2 [1])**.** *Consider a two armed bandit problem with reward distributions $\Theta = \{f_{R_1}(r), f_{R_2}(r)\}$, where the reward distribution of the optimal arm is $f_{R_1}(r)$ and for the sub-optimal arm is $f_{R_2}(r)$, and $\mathbb{E}\left[f_{R_1}(r)\right] > \mathbb{E}\left[f_{R_2}(r)\right]$; i.e., arm 1 is optimal. If it is possible to create an alternate problem with distributions $\Theta' = \{f_{R_1}(r), \tilde{f}_{R_2}(r)\}$ such that $\mathbb{E}\left[\tilde{f}_{R_2}(r)\right] > \mathbb{E}\left[f_{R_1}(r)\right]$ and $0 < D(f_{R_2}(r)||\tilde{f}_{R_2}(r)) < \infty$ (equivalent to assumption 1.6 in [1]), then for any policy that achieves sub-polynomial regret, we have*

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[n_2(T)\right]}{\log T} \geq \frac{1}{D(f_{R_2}(r)||\tilde{f}_{R_2}(r))}.$$

Proof. Proof of this is derived from the analysis done in [35]. We show the analysis here for completeness. A bandit instance v is defined by the reward distribution of arm 1 and arm 2. Since policy π achieves sub-polynomial regret, for any instance v, $\mathbb{E}_{v,\pi}\left[(Reg(T))\right] = \mathrm{O}(T^p)$ as $T \to \infty$, for all $p > 0$.

Consider the bandit instances $\Theta = \{f_{R_1}(r), f_{R_2}(r)\}$, $\Theta' = \{f_{R_1}(r), \tilde{f}_{R_2}(r)\}$, where $\mathbb{E}\left[f_{R_2}(r)\right] < \mathbb{E}\left[f_{R_1}(r)\right] < \mathbb{E}\left[\tilde{f}_{R_2}(r)\right]$. The bandit instance Θ' is constructed by changing the reward distribution of arm 2 in the original instance, in

such a way that arm 2 becomes optimal in instance Θ' without changing the reward distribution of arm 1 from the original instance.

From divergence decomposition lemma (derived in [35]), it follows that

$$D(\mathbb{P}_{\Theta,\Pi}||\mathbb{P}_{\Theta',\Pi}) = \mathbb{E}_{\Theta,\pi}\left[n_2(T)\right] D(f_{R_2}(r)||\tilde{f}_{R_2}(r)).$$

The high probability Pinsker's inequality (Lemma 2.6 from [36], originally in [37]) gives that for any event A,

$$\mathbb{P}_{\Theta,\pi}(A) + \mathbb{P}_{\Theta',\pi}(A^c) \geq \frac{1}{2}\exp\left(-D(\mathbb{P}_{\Theta,\pi}||\mathbb{P}_{\Theta',\pi})\right),$$

or equivalently,

$$D(\mathbb{P}_{\Theta,\pi}||\mathbb{P}_{\Theta',\pi}) \geq \log\frac{1}{2(\mathbb{P}_{\Theta,\pi}(A) + \mathbb{P}_{\Theta',\pi}(A^c))}.$$

If arm 2 is suboptimal in a 2-armed bandit problem, then $\mathbb{E}\left[Reg(T)\right] = \Delta_2\mathbb{E}\left[n_2(T)\right]$. Expected regret in Θ is

$$\mathbb{E}_{\Theta,\pi}\left[Reg(T)\right] \geq \frac{T\Delta_2}{2}\mathbb{P}_{\Theta,\pi}\left(n_2(T) \geq \frac{T}{2}\right),$$

Similarly regret in bandit instance Θ' is

$$\mathbb{E}_{\Theta',\pi}\left[Reg(T)\right] \geq \frac{T\delta}{2}\mathbb{P}_{\Theta',\pi}\left(n_2(T) < \frac{T}{2}\right),$$

since suboptimality gap of arm 1 in Θ' is δ. Define $\kappa(\Delta_2,\delta) = \frac{\min(\Delta_2,\delta)}{2}$. Then we have,

$$\mathbb{P}_{\Theta,\pi}\left(n_2(T) \geq \frac{T}{2}\right) + \mathbb{P}_{\Theta',\pi}\left(n_2(T) < \frac{T}{2}\right) \leq \frac{\mathbb{E}_{\Theta,\pi}\left[Reg(T)\right] + \mathbb{E}_{\Theta',\pi}\left[Reg(T)\right]}{\kappa(\Delta_2,\delta)T}. \tag{32}$$

On applying the high probability Pinsker's inequality and divergence decomposition lemma stated earlier, we get

$$D(f_{R_2}(r)||\tilde{f}_{R_2}(r))\mathbb{E}_{\Theta,\pi}\left[n_2(T)\right] \geq \log\left(\frac{\kappa(\Delta_2,\delta)T}{2(\mathbb{E}_{\Theta,\pi}\left[Reg(T)\right] + \mathbb{E}_{\Theta',\pi}\left[Reg(T)\right])}\right) \tag{33}$$

$$= \log\left(\frac{\kappa(\Delta_2,\delta)}{2}\right) + \log(T) - \log(\mathbb{E}_{\Theta,\pi}\left[Reg(T)\right] + \mathbb{E}_{\Theta',\pi}\left[Reg(T)\right]). \tag{34}$$

Since policy π achieves sub-polynomial regret for any bandit instance, $\mathbb{E}_{\Theta,\pi}\left[Reg(T)\right] + \mathbb{E}_{\Theta',\pi}\left[Reg(T)\right] \leq \gamma T^p$ for all T and any $p > 0$, hence,

$$\liminf_{T\to\infty} D(f_{R_2}(r)||\tilde{f}_{R_2}(r))\frac{\mathbb{E}_{\Theta,\pi}\left[n_2(T)\right]}{\log T} \geq 1 - \limsup_{T\to\infty}\frac{\mathbb{E}_{\Theta,\pi}\left[Reg(T)\right] + \mathbb{E}_{\Theta',\pi}\left[Reg(T)\right]}{\log T} + \liminf_{T\to\infty}\frac{\log\left(\frac{\kappa(\Delta_2,\delta)}{2}\right)}{\log T} \tag{35}$$

$$= 1. \tag{36}$$

Hence, $\liminf_{T\to\infty}\frac{\mathbb{E}_{\Theta,\pi}[n_2(T)]}{\log T} \geq \frac{1}{D(f_{R_2}(r)||\tilde{f}_{R_2}(r))}$. □

B. Results for any C-BANDIT Algorithm

Lemma 5. *Define $E_1(t)$ to be the event that arm k^* is empirically non-competitive in round $t+1$, then,*

$$\Pr(E_1(t)) \leq 2Kt\exp\left(\frac{-t\Delta_{min}^2}{2K}\right),$$

where $\Delta_{min} = \min_k \Delta_k$, the gap between the best and second-best arms.

Proof. The arm k^* is empirically non-competitive at round t if $k^* \neq k^{\text{emp}}$ and the empirical psuedo-reward of arm k^* with respect to arms $\ell \in \mathcal{S}_t$ is smaller than $\hat{\mu}_{k^{\text{emp}}}(t)$. This event can only occur if at-least one of the two following conditions is satisfied, i) the empirical mean of $k^{\text{emp}} \neq k^*$ is greater than $\mu_k^* - \frac{\Delta_{\min}}{2}$ or ii) the empirical pseudo-reward of arm k^* with respect to arms in $\mathcal{S}_t$ is smaller than $\mu_{k^*} - \frac{\Delta_{\min}}{2}$. We use this observation to analyse the $\Pr(E_1(t))$.

$$\Pr(E_1(t)) \leq \Pr\left(\left(\max_{\{\ell:n_\ell(t)>t/K,\ell\neq k^*\}}\hat{\mu}_\ell(t) > \mu_{k^*} - \frac{\Delta_{\min}}{2}\right) \bigcup \left(\min_{\{\ell:n_\ell(t)>t/K\}}\hat{\phi}_{k^*,\ell}(t) < \mu_{k^*} - \frac{\Delta_{\min}}{2}\right)\right) \tag{37}$$

$$\leq \Pr\left(\max_{\{\ell:n_\ell(t)>t/K,\ell\neq k^*\}}\hat{\mu}_\ell(t) > \mu_{k^*} - \frac{\Delta_{\min}}{2}\right) + \Pr\left(\min_{\{\ell:n_\ell(t)>t/K\}}\hat{\phi}_{k^*,\ell}(t) < \mu_{k^*} - \frac{\Delta_{\min}}{2}\right) \tag{38}$$

$$\leq \sum_{\ell\neq k^*}\Pr\left(\hat{\mu}_\ell(t) > \mu_{k^*} - \frac{\Delta_{\min}}{2}, n_\ell(t) > \frac{t}{K}\right) + \sum_{\ell=1}^{K}\Pr\left(\hat{\phi}_{k^*,\ell}(t) < \mu_{k^*} - \frac{\Delta_{\min}}{2}, n_\ell(t) > \frac{t}{K}\right) \tag{39}$$

$$= \sum_{\ell\neq k^*}\Pr\left(\hat{\mu}_\ell(t) - \mu_\ell > \mu_{k^*} - \mu_\ell - \frac{\Delta_{\min}}{2}, n_\ell(t) > \frac{t}{K}\right) + \sum_{\ell=1}^{K}\Pr\left(\hat{\phi}_{k^*,\ell}(t) - \phi_{k^*,\ell} < \mu_{k^*} - \phi_{k^*,\ell} - \frac{\Delta_{\min}}{2}, n_\ell(t) > \frac{t}{K}\right) \tag{40}$$

$$\leq \sum_{\ell\neq k^*}\Pr\left(\frac{\sum_{\tau=1}^{t}\mathbb{1}_{\{k_\tau=\ell\}}r_\tau}{n_\ell(t)} - \mu_\ell > \frac{\Delta_{\min}}{2}, n_\ell(t) > \frac{t}{K}\right) + \sum_{\ell=1}^{K}\Pr\left(\frac{\sum_{\tau=1}^{t}\mathbb{1}_{\{k_\tau=\ell\}}s_{k^*,\ell}(r_\tau)}{n_\ell(t)} - \phi_{k^*,\ell} < -\frac{\Delta_{\min}}{2}, n_\ell(t) > \frac{t}{K}\right) \tag{41}$$

$$\leq 2Kt\exp\left(\frac{-t\Delta_{\min}^2}{2K}\right), \tag{42}$$

Here (38) follows from union bound. We have (42) from the Hoeffding's inequality, as we note that rewards $\{r_\tau : \tau = 1,\ldots,t,\ k_\tau = k\}$ and pseudo-rewards $\{s_{k^*,l} : \tau_1,\ldots,t,\ k_\tau = l\}$ form a collection of i.i.d. random variables each of which is bounded between $[-1,1]$ with mean μ_k and $\phi_{k^*,l}$. The term

t before the exponent in (42) arises as the random variable $n_k(t)$ can take values from t/K to t (Lemma 1).

□

Lemma 6. *For a sub-optimal arm $k \neq k^*$ with sub-optimality gap Δ_k,*

$$\Pr\left(k = k^{emp}(t), n_{k^*}(t) \geq \frac{t}{K}\right) \leq t \exp\left(\frac{-t\Delta_k^2}{2K}\right).$$

Proof. We bound this probability as,

$$\begin{aligned}
&\Pr\left(k = k^{\text{emp}}(t), n_{k^*}(t) \geq \frac{t}{K}\right) \\
&= \Pr\left(k = k^{\text{emp}}(t), n_{k^*}(t) \geq \frac{t}{K}, n_k(t) \geq \frac{t}{K}\right) && (43)\\
&\leq \Pr\left(\hat{\mu}_k(t) \geq \hat{\mu}_{k^*}(t), n_k(t) \geq \frac{t}{K}, n_{k^*}(t) \geq \frac{t}{K}\right) && (44)\\
&\leq \Pr\left(\left(\hat{\mu}_{k^*}(t) < \mu_{k^*} - \frac{\Delta_k}{2} \bigcup \hat{\mu}_k(t) > \mu_{k^*} - \frac{\Delta_k}{2}\right),\right. \\
&\qquad \left. n_k(t) \geq \frac{t}{K}, n_{k^*}(t) \geq \frac{t}{K}\right) && (45)\\
&= \Pr\left(\left(\hat{\mu}_{k^*}(t) < \mu_{k^*} - \frac{\Delta_k}{2} \bigcup \hat{\mu}_k(t) > \mu_k + \frac{\Delta_k}{2}\right),\right. \\
&\qquad \left. n_k(t) \geq \frac{t}{K}, n_{k^*}(t) \geq \frac{t}{K}\right) && (46)\\
&\leq \Pr\left(\hat{\mu}_{k^*}(t) - \mu_{k^*} < -\frac{\Delta_k}{2}, n_{k^*}(t) \geq \frac{t}{K}\right) + \\
&\qquad \Pr\left(\hat{\mu}_k(t) - \mu_k > \frac{\Delta_k}{2}, n_k(t) \geq \frac{t}{K}\right) && (47)\\
&\leq 2t \exp\left(\frac{-t\Delta_k^2}{2K}\right) && (48)
\end{aligned}$$

We have (43) as arm k needs to be pulled at least $\frac{t}{K}$ in order to be arm $k^{\text{emp}}(t)$ at round t. The selection of k^{emp} is only done from the set of arms that have been pulled atleast $\frac{t}{K}$ times. Here, (48) follows from the Hoeffding's inequality. The term t before the exponent in (48) arises as the random variable $n_k(t)$ can take values from t/K to t (Lemma 1).

□

Lemma 7. *If for a suboptimal arm $k \neq k^*$, $\tilde{\Delta}_{k,k^*} > 0$, then,*

$$\Pr(k_{t+1} = k, n_{k^*}(t) = \max_k n_k(t)) \leq t \exp\left(\frac{-2t\tilde{\Delta}_{k,k^*}^2}{K}\right).$$

Moreover, if $\tilde{\Delta}_{k,k^*} \geq \sqrt{\frac{2K \log t_0}{t_0}}$ for some constant $t_0 > 0$. Then,

$$\Pr(k_{t+1} = k, n_{k^*}(t) = \max_k n_k(t)) \leq t^{-3} \quad \forall t > t_0.$$

Proof. We now bound this probability as,

$$\begin{aligned}
&\Pr(k_{t+1} = k, n_{k^*}(t) = \max_k n_k(t)) \\
&\leq \Pr\left(k_{t+1} = k, n_{k^*}(t) \geq \frac{t}{K}\right) && (49)\\
&= \Pr\left(k \in \{\mathcal{A}_t \cup \{k_{\text{emp}}(t)\}\}, k_{t+1} = k, n_{k^*}(t) \geq \frac{t}{K}\right) && (50)\\
&\leq \Pr\left(k \in \mathcal{A}_t, k_{t+1} = k, n_{k^*}(t) \geq \frac{t}{K}\right) + \\
&\qquad \Pr\left(k = k^{\text{emp}}(t), n_{k^*}(t) \geq \frac{t}{K}\right) && (51)\\
&\leq \Pr\left(k \in \mathcal{A}_t, k_{t+1} = k, n_{k^*}(t) \geq \frac{t}{K}\right) + 2t \exp\left(\frac{-t\Delta_k^2}{2K}\right) && (52)\\
&\leq \Pr\left(\hat{\mu}_{k^*}(t) < \hat{\phi}_{k,k^*}(t), k_{t+1} = k, n_{k^*}(t) \geq \frac{t}{K}\right) + \\
&\qquad 2t \exp\left(\frac{-t\Delta_k^2}{2K}\right) && (53)\\
&\leq \Pr\left(\hat{\mu}_{k^*}(t) < \hat{\phi}_{k,k^*}(t), n_{k^*}(t) \geq \frac{t}{K}\right) + 2t \exp\left(\frac{-t\Delta_k^2}{2K}\right) && (54)\\
&\leq \Pr\left(\frac{\sum_{\tau=1}^t \mathbb{1}_{\{k_\tau = k^*\}} r_\tau}{n_{k^*}(t)} < \frac{\sum_{\tau=1}^t \mathbb{1}_{\{k_\tau = k^*\}} s_{k,k^*}(r_\tau)}{n_{k^*(t)}},\right. \\
&\qquad \left. n_{k^*}(t) \geq \frac{t}{K}\right) + 2t \exp\left(\frac{-t\Delta_k^2}{2K}\right) && (55)\\
&= \Pr\left(\frac{\sum_{\tau=1}^t \mathbb{1}_{\{k_\tau = k^*\}}(r_\tau - s_{k,k^*})}{n_{k^*}(t)} - (\mu_{k^*} - \phi_{k,k^*}) < \right. \\
&\qquad \left. - \tilde{\Delta}_{k,k^*}, n_{k^*} \geq \frac{t}{K}\right) + 2t \exp\left(\frac{-t\Delta_k^2}{2K}\right) && (56)\\
&\leq t \exp\left(\frac{-t\tilde{\Delta}_{k,k^*}^2}{2K}\right) + 2t \exp\left(\frac{-t\Delta_k^2}{2K}\right) && (57)\\
&\leq 3t^{-3} \quad \forall t > t_0. && (58)
\end{aligned}$$

We have (49) as $n_{k^*}(t)$ needs to be at-least $\frac{t}{K}$ for $n_{k^*}(t)$ to be $\max_k n_k(t)$. Equation (50) holds as arm k needs to be in the set $\{\mathcal{A}_t \cup \{k^{\text{emp}}(t)\}\}$ to be selected by C-BANDIT at round t. Inequality (52) arises from the result of Lemma 6. The inequality (53) follows as $\phi_{k,k^*} > \hat{\mu}_{k^*}$ is a necessary condition for arm k to be in the competitive set $\mathcal{A}_t$ at round t. Here, (56) follows from the Hoeffding's inequality as we note that rewards $\{r_\tau - s_{k,k^*}(r_\tau) : \tau = 1, \ldots, t,\ k_\tau = k^*\}$ form a collection of i.i.d. random variables each of which is bounded between $[-1, 1]$ with mean $(\mu_k - \phi_{k,k^*})$. The term t before the exponent in (56) arises as the random variable $n_k(t)$ can take values from t/K to t (Lemma 1). Step (58) follows from the fact that $\tilde{\Delta}_{k,k^*} \geq 2\sqrt{\frac{2K \log t_0}{t_0}}$ for some constant $t_0 > 0$. □

C. Algorithm specific results for C-UCB

Lemma 8. *If* $\Delta_{min} \geq 4\sqrt{\frac{2K\log t_0}{t_0}}$ *for some constant* $t_0 > 0$, *then,*

$$\Pr(k_{t+1} = k, n_k(t) \geq s) \leq 3t^{-3} \quad \text{for } s > \frac{t}{2K}, \forall t > t_0.$$

Proof. By noting that $k_{t+1} = k$ corresponds to arm k having the highest index among the set of arms that are not empirically *non-competitive* (denoted by $\mathcal{A}$), we have,

$$\Pr(k_{t+1} = k, n_k(t) \geq s) =$$
$$\Pr(I_k(t) = \arg\max_{k' \in \mathcal{A}} I_{k'}(t), n_k(t) \geq s) \tag{59}$$
$$\leq \Pr(E_1(t) \cup (E_1^c(t), I_k(t) > I_{k^*}(t)), n_k(t) \geq s) \tag{60}$$
$$\leq \Pr(E_1(t), n_k(t) \geq s) + \Pr(E_1^c(t), I_k(t) > I_{k^*}(t), n_k(t) \geq s) \tag{61}$$
$$\leq 2Kt\exp\left(\frac{-t\Delta_{\min}^2}{2K}\right) + \Pr\left(I_k(t) > I_{k^*}(t), n_k(t) \geq s\right). \tag{62}$$

Here $E_1(t)$ is the event described in Lemma 5. If arm k^* is not empirically non-competitive at round t, then arm k can only be pulled in round $t+1$ if $I_k(t) > I_{k^*}(t)$, due to which we have (60). Inequalities (61) and (62) follow from union bound and Lemma 5 respectively.

We now bound the second term in (62).

$$\Pr(I_k(t) > I_{k^*}(t), n_k(t) \geq s) =$$
$$\Pr\left(I_k(t) > I_{k^*}(t), n_k(t) \geq s, \mu_{k^*} \leq I_{k^*}(t)\right) + \Pr\left(I_k(t) > I_{k^*}(t), n_k(t) \geq s | \mu_{k^*} > I_{k^*}(t)\right) \times \Pr\left(\mu_{k^*} > I_{k^*}(t)\right) \tag{63}$$
$$\leq \Pr\left(I_k(t) > I_{k^*}(t), n_k(t) \geq s, \mu_{k^*} \leq I_{k^*}(t)\right) + \Pr\left(\mu_{k^*} > I_{k^*}(t)\right) \tag{64}$$
$$\leq \Pr\left(I_k(t) > I_{k^*}(t), n_k(t) \geq s, \mu_{k^*} \leq I_{k^*}(t)\right) + t^{-3} \tag{65}$$
$$= \Pr\left(I_k(t) > \mu_{k^*}, n_k(t) \geq s\right) + t^{-4} \tag{66}$$
$$= \Pr\left(\hat{\mu}_k(t) + \sqrt{\frac{2\log t}{n_k(t)}} > \mu_{k^*}, n_k(t) \geq s\right) + t^{-3} \tag{67}$$
$$= \Pr\left(\hat{\mu}_k(t) - \mu_k > \mu_{k^*} - \mu_k - \sqrt{\frac{2\log t}{n_k(t)}}, n_k(t) \geq s\right) + t^{-3} \tag{68}$$
$$= \Pr\left(\frac{\sum_{\tau=1}^{t} \mathbb{1}_{\{k_\tau = k\}} r_\tau}{n_k(t)} - \mu_k > \Delta_k - \sqrt{\frac{2\log t}{n_k(t)}}, n_k(t) \geq s\right) + t^{-3} \tag{69}$$
$$\leq t\exp\left(-2s\left(\Delta_k - \sqrt{\frac{2\log t}{s}}\right)^2\right) + t^{-3} \tag{70}$$
$$\leq t^{-3}\exp\left(-2s\left(\Delta_k^2 - 2\Delta_k\sqrt{\frac{2\log t}{s}}\right)\right) + t^{-3} \tag{71}$$
$$\leq 2t^{-3} \quad \text{for all } t > t_0. \tag{72}$$

We have (63) holds because of the fact that $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$, Inequality (65) follows from Lemma 2. From the definition of $I_k(t)$ we have (67). Inequality (70) follows from Hoeffding's inequality and the term t before the exponent in (70) arises as the random variable $n_k(t)$ can take values from s to t (Lemma 1). Inequality (72) follows from the fact that $s > \frac{t}{2K}$ and $\Delta_k \geq 4\sqrt{\frac{2K\log t_0}{t_0}}$ for some constant $t_0 > 0$.

Plugging this in the expression of $\Pr(k_t = k, n_k(t) \geq s)$ (62) gives us,

$$\Pr(k_{t+1} = k, n_k(t) \geq s) \leq$$
$$2Kt\exp\left(\frac{-t\Delta_{\min}^2}{2K}\right) + \Pr(I_k(t) > I_{k^*}(t), n_k(t) \geq s) \tag{73}$$
$$\leq 2Kt\exp\left(\frac{-t\Delta_{\min}^2}{2K}\right) + 2t^{-3} \tag{74}$$
$$\leq 2(K+1)t^{-3}. \tag{75}$$

Here, (75) follows from the fact that $\Delta_{\min} \geq 4\sqrt{\frac{2K\log t_0}{t_0}}$ for some constant $t_0 > 0$. □

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

$$\Pr\left(n_k(t) > \frac{t}{K}\right) \leq (2K+2)K\left(\frac{t}{K}\right)^{-2} \quad \forall t > Kt_0.$$

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

$$\Pr\left(n_k(t) \geq \frac{t}{K}\right) =$$
$$\Pr\left(n_k(t) \geq \frac{t}{K} \mid n_k(t-1) \geq \frac{t}{K}\right)\Pr\left(n_k(t-1) \geq \frac{t}{K}\right) + \Pr\left(k_t = k, n_k(t-1) = \frac{t}{K} - 1\right) \tag{76}$$
$$\leq \Pr\left(n_k(t-1) \geq \frac{t}{K}\right) + \Pr\left(k_t = k, n_k(t-1) = \frac{t}{K} - 1\right) \tag{77}$$
$$\leq \Pr\left(n_k(t-1) \geq \frac{t}{K}\right) + (2K+2)(t-1)^{-3}, \quad \forall (t-1) > t_0. \tag{78}$$

Here, (78) follows from Lemma 8.

This gives us

$$\Pr\left(n_k(t) \geq \frac{t}{K}\right) - \Pr\left(n_k(t-1) \geq \frac{t}{K}\right) \leq (2K+2)(t-1)^{-3}, \quad \forall (t-1) > t_0.$$

Now consider the summation

$$\sum_{\tau=\frac{t}{K}}^{t} \Pr\left(n_k(\tau) \geq \frac{t}{K}\right) - \Pr\left(n_k(\tau-1) \geq \frac{t}{K}\right) \leq \sum_{\tau=\frac{t}{K}}^{t} (2K+2)(\tau-1)^{-3}.$$

This gives us,

$$\Pr\left(n_k(t) \geq \frac{t}{K}\right) - \Pr\left(n_k\left(\frac{t}{K}-1\right) \geq \frac{t}{K}\right) \leq \sum_{\tau=\frac{t}{K}}^{t} (2K+2)(\tau-1)^{-3}.$$

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

$$\Pr\left(n_k(t) \geq \frac{t}{K}\right) \leq \sum_{\tau=\frac{t}{K}}^{t} (2K+2)(\tau-1)^{-3} \tag{79}$$

$$\leq (2K+2)K\left(\frac{t}{K}\right)^{-2} \quad \forall t > Kt_0. \tag{80}$$

□

D. Regret Bounds for C-UCB

Proof of Theorem 1 We bound $\mathbb{E}[n_k(T)]$ as,

$$\mathbb{E}[n_k(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{k_t=k\}}\right] \tag{81}$$

$$= \sum_{t=0}^{T-1} \Pr(k_{t+1}=k) \tag{82}$$

$$= \sum_{t=1}^{Kt_0} \Pr(k_t=k) + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1}=k) \tag{83}$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1}=k, n_{k^*}(t) = \max_{k'} n_{k'}(t)) + \sum_{t=Kt_0}^{T-1} \sum_{k'\neq k^*} \Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \times \Pr(k_{t+1}=k | n_{k'}(t) = \max_{k''} n_{k''}(t)) \tag{84}$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} \Pr(k_{t+1}=k, n_{k^*}(t) = \max_{k'} n_{k'}(t)) + \sum_{t=Kt_0}^{T-1} \sum_{k'\neq k^*} \Pr(n_{k'}(t) = \max_{k''} n_{k''}(t)) \tag{85}$$

$$\leq Kt_0 + \sum_{t=Kt_0}^{T-1} 3t^{-3} + \sum_{t=Kt_0}^{T} \sum_{k'\neq k^*} \Pr\left(n_{k'}(t) \geq \frac{t}{K}\right) \tag{86}$$

$$\leq Kt_0 + \sum_{t=1}^{T} 3t^{-3} + (K+1)K(K-1) \sum_{t=Kt_0}^{T} 2\left(\frac{t}{K}\right)^{-2}. \tag{87}$$

Here, (86) follows from Lemma 7 and (87) follows from Lemma 9.

Proof of Theorem 2

For any suboptimal arm $k \neq k^*$,

$$\mathbb{E}[n_k(T)] \leq \sum_{t=1}^{T} \Pr(k_t=k) \tag{88}$$

$$= \sum_{t=1}^{T} \Pr(E_1(t), k_t = k \cup (E_1^c(t), I_k > I_{k^*}), k_t = k) \tag{89}$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(E_1^c(t), I_k(t-1) > I_{k^*}(t-1), k_t = k) \tag{90}$$

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

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \Pr(I_k(t-1) > I_{k^*}(t-1), k_t = k) \tag{91}$$

$$= \sum_{t=1}^{T} 2Kt\exp\left(-\frac{t\Delta_{\min}^2}{2K}\right) + \sum_{t=0}^{T-1} \Pr\left(I_k(t) > I_{k^*}(t), k_t = k\right) \tag{92}$$

$$= \sum_{t=1}^{T} 2Kt\exp\left(-\frac{t\Delta_{\min}^2}{2K}\right) + \mathbb{E}\left[\mathbb{1}_{I_k > I_{k^*}}(T)\right] \tag{93}$$

$$\leq 8\frac{\log(T)}{\Delta_k^2} + \left(1+\frac{\pi^2}{3}\right) + \sum_{t=1}^{T} 2Kt\exp\left(-\frac{t\Delta_{\min}^2}{2K}\right). \tag{94}$$

Here, (92) follows from Lemma 5. We have (93) from the definition of $\mathbb{E}[n_{I_k > I_{k^*}}(T)]$ in Lemma 3, and (94) follows from Lemma 3.

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

E. Regret analysis for the C-TS Algorithm

We now present results for C-TS in the scenario where $K = 2$ and Thompson sampling is employed with Beta priors [29]. In order to prove results for C-TS, we assume that rewards are either 0 or 1. The Thompson sampling algorithm with beta prior, maintains a posterior distribution on mean of arm k as $Beta\left(n_k(t) \times \hat{\mu}_k(t) + 1, n_k(t) \times (1 - \hat{\mu}_k(t)) + 1\right)$. Subsequently, it generates a sample $S_k(t) \sim Beta\left(n_k(t) \times \hat{\mu}_k(t) + 1, n_k(t) \times (1 - \hat{\mu}_k(t)) + 1\right)$ for each arm k and selects the arm $k_{t+1} = \arg\max_{k\in\mathcal{K}} S_k(t)$. The C-TS algorithm with Beta prior uses this Thompson sampling procedure in its last step, i.e., $k_{t+1} = \arg\max_{k\in\mathcal{C}_t} S_k(t)$, where $\mathcal{C}_t$ is the set of empirically competitive arms at round t. We show that in a 2-armed bandit problem, the regret is $\mathrm{O}(1)$ if the sub-optimal arm k is non-competitive and is $\mathrm{O}(\log T)$ otherwise.

For the purpose of regret analysis of C-TS, we define two thresholds, a lower threshold L_k, and an upper threshold U_k for arm $k \neq k^*$,

$$U_k = \mu_k + \frac{\Delta_k}{3}, \qquad L_k = \mu_{k^*} - \frac{\Delta_k}{3}. \tag{95}$$

Let $E_i^{\mu}(t)$ and $E_i^{S}(t)$ be the events that,

$$E_k^{\mu}(t) = \{\hat{\mu}_k(t) \leq U_k\}$$
$$E_k^{S}(t) = \{S_k(t) \leq L_k\}. \tag{96}$$

To analyse the regret of C-TS, we first show that the number of times arm k is pulled jointly with the event that $n_k(t-1) \geq \frac{t}{2}$ is bounded above by an $\mathrm{O}(1)$ constant, which is independent of the total number of rounds T.

Lemma 10. *If $\Delta_k \geq 4\sqrt{\frac{2K\log t_0}{t_0}}$ for some constant $t_0 > 0$, then,*

$$\sum_{t=2t_0}^{T} \Pr\left(k_t = k, n_k(t-1) \geq \frac{t}{2}\right) = \mathrm{O}(1)$$

where $k \neq k^*$ is a sub-optimal arm.

Proof. We start by bounding the probability of the pull of k-th arm at round t as follows,

$$\begin{aligned}
&\Pr\left(k_t = k, n_k(t-1) \geq \frac{t}{2}\right) \overset{(a)}{\leq} \\
&\quad \Pr\left(E_1(t), k_t = k, n_k(t-1) \geq \frac{t}{2}\right) + \\
&\quad \Pr\left(\overline{E_1(t)}, k_t = k, n_k(t-1) \geq \frac{t}{2}\right) \\
&\overset{(b)}{\leq} 2Kt\exp\left(\frac{-t\Delta_{\min}^2}{2K}\right) + \Pr\left(\overline{E_1(t)}, k_t = k, n_k(t-1) \geq \frac{t}{2}\right) \\
&\overset{(c)}{\leq} 2Kt^{-3} + \underbrace{\Pr\left(k_t = k, E_k^{\mu}(t), E_k^{S}(t), n_k(t-1) \geq \frac{t}{2}\right)}_{\textbf{term A}} + \\
&\underbrace{\Pr\left(k_t = k, E_k^{\mu}(t), \overline{E_k^{S}(t)}, n_k(t-1) \geq \frac{t}{2}\right)}_{\textbf{term B}} + \\
&\underbrace{\Pr\left(k_t = k, \overline{E_k^{\mu}(t)}, n_k(t-1) \geq \frac{t}{2}\right)}_{\textbf{term C}}
\end{aligned} \tag{97}$$

where (b), comes from Lemma 5. Here, (97) follows from the fact that $\Delta_{\min} \geq 4\sqrt{\frac{2K\log t_0}{t_0}}$ for some constant $t_0 > 0$. Now we treat each term in (97) individually. To bound term A, we note that $\Pr\left(k_t = k, E_k^{\mu}(t), E_k^{S}(t), n_k(t-1) \geq \frac{t}{2}\right) \leq \Pr\left(k_t = k, E_k^{\mu}(t), E_k^{S}(t)\right)$. From the analysis in [29] (equation 6), we see that $\sum_{t=1}^{T}\Pr\left(k_t = k, E_k^{\mu}(t), E_k^{S}(t)\right) = \mathrm{O}(1)$ as it is shown through Lemma 2 in [29] that,
$\sum_{t=1}^{T}\Pr\left(k_t = k, E_k^{\mu}(t), E_k^{S}(t)\right) \leq \frac{216}{\Delta_k^2} + \sum_{j=0}^{T}\Theta\left(e^{-\frac{\Delta_k^2 j}{18}} + \frac{1}{e^{\frac{\Delta_k^2 j}{36}}-1} + \frac{9}{(j+1)\Delta_k^2}e^{-D_k j}\right)$.
Here, $D_k = L_k\log\frac{L_k}{\mu_{k^*}} + (1-L_k)\log\frac{1-L_k}{1-\mu_{k^*}}$. Due to this, $\sum_{t=2t_0}^{T}\Pr\left(k_t = k, E_k^{\mu}(t), E_k^{S}(t), n_k(t-1) \geq \frac{t}{2}\right) = \mathrm{O}(1)$.

We now bound the sum of term B from $t = 1$ to T by noting that
$\Pr\left(k_t = k, E_k^{\mu}(t), \overline{E_k^{S}(t)}, n_k(t-1) \geq \frac{t}{2}\right) \leq \Pr\left(k_t = k, \overline{E_k^{S}(t)}\right)$. Additionally, from Lemma 3 in [29], we get that $\sum_{t=1}^{T}\Pr\left(k_t = k, \overline{E_k^{S}(t)}\right) \leq \frac{1}{d(U_k,\mu_k)} + 1$, where $d(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$. As a result, we see that $\sum_{t=1}^{T}\Pr\left(k_t = k, E_k^{\mu}(t), \overline{E_k^{S}(t)}, n_k(t-1) \geq \frac{t}{2}\right) = \mathrm{O}(1)$.

Finally, for the last term C we can show that,

$$\begin{aligned}
(C) &= \Pr\left(k_t = k, \overline{E_k^{\mu}(t)}, n_k(t-1) \geq \frac{t}{2}\right) \\
&\leq \Pr\left(\overline{E_k^{\mu}(t)}, n_k(t-1) \geq \frac{t}{2}\right) \\
&= \Pr\left(\hat{\mu}_k - \mu_k > \frac{\Delta_k}{3}, n_k(t-1) \geq \frac{t}{2}\right) \\
&\leq t\exp\left(-2\frac{t}{2}\frac{\Delta_k^2}{9}\right) \qquad (98) \\
&\leq t^{-3}
\end{aligned}$$

Here (98) follows from hoeffding's inequality and the union bound trick to handle random variable $n_k(t-1)$. After plugging these results in (97), we get that

$$\begin{aligned}
&\sum_{t=2t_0}^{T}\Pr\left(k_t = k, n_k(t-1) \geq \frac{t}{2}\right) \leq \\
&\quad \sum_{t=2t_0}^{T} 2Kt^{-3} + \\
&\quad \sum_{t=2t_0}^{T}\Pr\left(k_t = k, E_k^{\mu}(t), E_k^{S}(t), n_k(t-1) \geq \frac{t}{2}\right) + \\
&\quad \sum_{t=2t_0}^{T}\Pr\left(k_t = k, E_k^{\mu}(t), \overline{E_k^{S}(t)}, n_k(t-1) \geq \frac{t}{2}\right) + \\
&\quad \sum_{t=2t_0}^{T}\Pr\left(k_t = k, \overline{E_k^{\mu}(t)}, n_k(t-1) \geq \frac{t}{2}\right) \qquad (99) \\
&\leq \sum_{t=2t_0}^{T} 2Kt^{-3} + \mathrm{O}(1) + \mathrm{O}(1) + \sum_{t=2t_0}^{T} t^{-3} \qquad (100) \\
&= \mathrm{O}(1) \qquad (101)
\end{aligned}$$

□

We now show that the expected number of pulls by C-TS for a non-competitive arm is bounded above by an $\mathrm{O}(1)$ constant.

Expected number of pulls by C-TS for a non-competitive arm. We bound $\mathbb{E}\left[n_k(t)\right]$ as

$$\begin{aligned}
\mathbb{E}\left[n_k(T)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{k_t=k\}}\right] \\
&= \sum_{t=0}^{T-1}\Pr(k_{t+1} = k) \\
&= \sum_{t=1}^{2t_0}\Pr(k_t = k) + \sum_{t=2t_0}^{T-1}\Pr(k_{t+1} = k) \\
&\leq 2t_0 + \sum_{t=2t_0}^{T-1}\Pr\left(k_{t+1} = k, n_{k^*}(t) \geq \frac{t}{2}\right) + \\
&\quad \sum_{t=2t_0}^{T-1}\Pr\left(k_{t+1} = k, n_k(t) \geq \frac{t}{2}\right) \qquad (102)
\end{aligned}$$

$$\leq 2t_0 + \sum_{t=2t_0}^{T-1} 3t^{-3} + \sum_{t=2t_0}^{T-1} \Pr\left(k_{t+1} = k, n_k(t) \geq \frac{t}{2}\right) \tag{103}$$

$$= \mathrm{O}(1) \tag{104}$$

Here, (103) follows from Lemma 7 and (104) follows from Lemma 10 and the fact that the sum of $3t^{-3}$ is bounded and $t_0 = \inf\left\{\tau > 0 : \Delta_{\min}, \epsilon_k \geq 4\sqrt{\frac{2K\log\tau}{\tau}}\right\}$.

We now show that when the sub-optimal arm k is competitive, the expected pulls of arm k is $\mathrm{O}(\log T)$.

Expected number of pulls by C-TS for a competitive arm $k \neq k^*$.: For any sub-optimal arm $k \neq k^*$,

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

$$= \sum_{t=1}^{T} \Pr((k_t = k, E_1(t)) \cup (E_1^c(t), k_t = k)) \tag{105}$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \sum_{t=1}^{T} \Pr(E_1^c(t), k_t = k)$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \sum_{t=1}^{T} \Pr(E_1^c(t), k_t = k, S_k(t-1) > S_{k^*}(t-1)) \tag{106}$$

$$\leq \sum_{t=1}^{T} \Pr(E_1(t)) + \sum_{t=0}^{T-1} \Pr(S_k(t) > S_{k^*}(t), k_{t+1} = k)$$

$$= \sum_{t=1}^{T} 2Kt^{-3} + \sum_{t=0}^{T-1} \Pr\left(S_k(t) > S_{k^*}(t), k_{t+1} = k\right) \tag{107}$$

$$\leq \frac{9\log(T)}{\Delta_k^2} + \mathrm{O}(1) + \sum_{t=1}^{T} 2Kt^{-3}. \tag{108}$$

$$= \mathrm{O}(\log T). \tag{109}$$

Here, (107) follows from Lemma 5. We have (108) from the analysis of Thompson Sampling for the classical bandit problem in [29]. This arises as the term $\Pr\left(S_k(t) > S_{k^*}(t), k_{t+1} = k\right)$ counts the number of times $S_k(t) > S_{k^*}(t)$ and $k_{t+1} = k$. This is precisely the term analysed in Theorem 3 of [29] to bound the expected pulls of sub-optimal arms by TS. In particular, [29] analyzes the expected number of pull of sub-optimal arm (termed as $\mathbb{E}\left[k_i(T)\right]$ in their paper) by evaluating $\sum_{t=0}^{T-1} \Pr(S_k(t) > S_{k^*}(t), k_{t+1} = k)$ and it is shown in their Section 2.1 (proof of Theorem 1 of [29]) that $\sum_{t=0}^{T-1} \Pr(S_k(t) > S_{k^*}(t), k_{t+1} = k) \leq \mathrm{O}(1) + \frac{\log(T)}{d(x_i, y_i)}$. The term x_i is equivalent to U_k and y_i is equal to L_k in our notations. Moreover $d(U_k, L_k) \leq \frac{\Delta_k^2}{9}$, giving us the desired result of (108).

F. Lower Bounds

For the proof we define $R_k = Y_k(X)$ and $\tilde{R}_k = g_k(\tilde{X})$, where $f_X(x)$ is the probability density function of random variable X and $f_{\tilde{X}}(x)$ is the probability density function of random variable $\tilde{X}$. Similarly, we define $f_{R_k}(r)$ to be the reward distribution of arm k.

Proof of Theorem 4

Let arm k be a *Competitive* sub-optimal arm, i.e $\tilde{\Delta}_{k,k^*} < 0$. To prove that regret is $\Omega(\log T)$ in this setting, we need to create a new bandit instance, in which reward distribution of optimal arm is unaffected, but a previously competitive sub-optimal arm k becomes optimal in the new environment. We do so by constructing a bandit instance with latent randomness $\tilde{X}$ and random rewards $\tilde{Y}_k(X)$. Let's denote to $\tilde{Y}_k(\tilde{X})$ to be the random reward obtained on pulling arm k given the realization of $\tilde{X}$. To make arm k optimal in the new bandit instance, we construct $\tilde{Y}_k(X)$ and $\tilde{X}$ in the following manner. Let $\mathcal{Y}_k$ denote the support of $Y_k(X)$.

Define

$$\tilde{Y}_k(X) = \begin{cases} \bar{g}_k(X) & \text{w.p. } 1-\epsilon_1 \\ \tilde{Y}_k(X) \sim \text{Uniform}(\mathcal{Y}_k) & \text{w.p. } \epsilon_1 \end{cases}$$

This changes the conditional reward of arm k in the new bandit instance (with increased mean).

Furthermore, Define

$$\tilde{X} = \begin{cases} S(R_{k^*}) & w.p. 1-\epsilon_2 \\ \text{Uniform} \sim \mathcal{X} & w.p. \epsilon_2. \end{cases},$$

with $S(R_{k^*}) = \arg\max_{\bar{g}_{k^*}(x) < R_{k^*} < \bar{g}_{k^*}(x)} \bar{g}_k(x)$.
Here R_{k^*} represents the random reward of arm k^* in the original bandit instance.

This construction of $\tilde{X}$ is possible for some $\epsilon_1, \epsilon_2 > 0$, whenever arm k is competitive by definition. Moreover, under such a construction one can change reward distribution of $\tilde{Y}_{k^*}(\tilde{X})$ such that reward $\tilde{R}_{k^*}$ has the same distribution as R_{k^*}. This is done by changing the conditional reward distribution, $f_{\tilde{Y}_{k^*}|X}(r) = \frac{f_{Y_{k^*}|X}(r) f_X(x)}{f_{\tilde{X}}(x)}$.

Due to this, if an arm is competitive, there exists a new bandit instance with latent randomness $\tilde{X}$ and conditional rewards $\tilde{Y}_{k^*}|X$ and $\tilde{Y}_k|X$ such that $f_{R_{k^*}} = f_{\tilde{R}_k^*}$ and $\mathbb{E}\left[\tilde{R}_k\right] > \mu_{k^*}$, with f_{R_k} denoting the probability distribution function of the reward from arm k and $\tilde{R}_k$ representing the reward from arm k in the new bandit instance.

Therefore, if these are the only two arms in our problem, then from Lemma 4,

$$\lim_{T\to\infty} \inf \frac{\mathbb{E}\left[n_k(T)\right]}{\log T} \geq \frac{1}{D(f_{R_k}(r)||f_{\tilde{R}_k}(r))},$$

where $f_{\tilde{R}_k}(r)$ represents the reward distribution of arm k in the new bandit instance.

Moreover, if we have more $K-1$ sub-optimal arms, instead of just 1, then

$$\lim_{T\to\infty} \inf \frac{\mathbb{E}\left[\sum_{\ell\neq k^*} n_\ell(T)\right]}{\log T} \geq \frac{1}{D(f_{R_k}(r)||f_{\tilde{R}_k}(r))}.$$

Consequently, since $\mathbb{E}\left[Reg(T)\right] = \sum_{ell=1}^{K} \Delta_\ell \mathbb{E}\left[n_\ell(T)\right]$, we have

$$\lim_{T\to\infty} \inf \frac{\mathbb{E}\left[Reg(T)\right]}{\log(T)} \geq \max_{k\in\mathcal{C}} \frac{\Delta_k}{D(f_{R_k}||f_{\tilde{R}_k})}. \tag{110}$$

r	$s_{2,1}(r)$
0	$\frac{2}{3}$
1	$\frac{6}{7}$

r	$s_{1,2}(r)$
0	$\frac{3}{4}$
1	$\frac{2}{3}$

(a)	$R_2=0$	$R_2=1$
$R_1=0$	0.1	0.2
$R_1=1$	0.3	0.4

(b)	$R_2=0$	$R_2=1$
$R_1=0$	a	b
$R_1=1$	c	d

TABLE V: The top row shows the pseudo-rewards of arms 1 and 2, i.e., upper bounds on the conditional expected rewards (which are known to the player). The bottom row depicts two possible joint probability distribution (unknown to the player). Under distribution (a), Arm 1 is optimal and all pseudo-reward except $s_{2,1}(1)$ are tight.

A stronger lower bound valid for the general case

A stronger lower bound for the general case can be shown by using the result in Proposition 1 of [38]. If $\mathcal{P}$ denotes the set of all possible joint probability distribution under which all pseudo-reward constraints are satisfied and P denotes the underlying unknown joint probability distribution which has k^* as the optimal arm. Then, the expected cumulative regret for any algorithm that achieves a sub-polynomial regret is lower bounded as

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

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

$$\begin{aligned} &\min_{\eta(k)\geq 0, k\in\mathcal{K}} \sum_{k\in\mathcal{K}} \eta(k)\left(\max_{\ell\in\mathcal{K}} \mu_\ell - \mu_k\right) \\ &\text{subject to} \sum_{k\in\mathcal{K}} \eta(k) D(P,Q,k) \geq 1, \quad \forall Q\in\mathcal{Q}, \\ &\textit{where} \quad \mathcal{Q} = \{Q\in\mathcal{P} : f_R(R_{k^*}|Q,k^*) = f_R(R_{k^*}|P,k^*) \\ &\qquad \text{and} \quad k^* \neq \arg\max_{k\in\mathcal{K}} \mu_k(Q)\}. \end{aligned} \tag{111}$$

Here, $D(P,Q,k)$ is the KL-Divergence between reward distributions of arm k under joint probability distributions P and Q, i.e., $f_R(R_k|\theta,k)$ and $f_R(R_k|\lambda,k)$. The term $\mu_k(Q)$ represents the mean reward of arm k under the joint probability distribution Q.

To interpret the lower bound, one can think of $\mathcal{Q}$ as the set of all joint probability distributions, under which the reward distribution of arm k^* remains the same, but some other arm $k'\neq k^*$ is optimal under the joint probability distribution. The optimization problem reflects the amount of samples needed to distinguish these two joint probability distributions. This result is based on the original result of [39], which has been used recently in [13], [38] for studying other bandit problems.

Lower bound discussion in general framework

Consider the example shown in Table V, for the joint probability distribution (a), Arm 1 is optimal. Moreover, all pseudo-rewards except $s_{2,1}(1)$ are tight, i.e.,$s_{\ell,k}(r) = \mathbb{E}\left[R_\ell | R_k = r\right]$. For the joint probability distribution shown in (a), expected pseudo-reward of Arm 2 is 0.8 and hence it is competitive. Due to this, our C-UCB and C-TS algorithms pull Arm 2 $\mathrm{O}(\log T)$ times.

However, it is not possible to construct an alternate bandit environment with joint probability distribution shown in Table V(b), such that Arm 2 becomes optimal while maintaining the same marginal distribution for Arm 1, and making sure that the pseudo-rewards still remain upper bound on conditional expected rewards. Formally, there does not exist a,b,c,d such that $c+d=0.7$, $\frac{c}{a+c} < 3/4$, $\frac{b}{a+b} < 2/3$, $\frac{d}{b+d} < 2/3$, $\frac{d}{d+c} < 6/7$ and $a+b+c+d=1$. This suggests that there should be a way to achieve $\mathrm{O}(1)$ regret in this scenario. We believe this can be done by using all the constraints (imposed by the knowledge of pair-wise pseudo-rewards to shrink the space of possible joint probability distributions) when calculating empirical pseudo-reward. However, this becomes tough to implement as the ratings can have multiple possible values and the number of arms is more than 2. We leave the task of coming up with a practically feasible and easy to implement algorithm that achieves bounded regret whenever possible in a general setup as an interesting open problem.

Samarth Gupta received the B.Tech. and M.Tech. in Electrical Engineering from Indian Institute of Technology, Bombay in 2017. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering at Carnegie Mellon University. His research focuses on online learning and statistical inference for applications in experiment design, A/B testing and recommendation systems.

Shreyas Chaudhari (S'20-S'21) received the B.Tech. degree majoring in Electrical Engineering with a minor in Machine Learning from Indian Institute of Technology Madras (India) in 2019, and the M.S. degree in Electrical and Computer Engineering from Carnegie Mellon University (US) in 2020. His research focuses on problems in reinforcement learning and multi-armed bandits, using tools from applied probability, machine learning, and deep learning. He is a member of the IEEE-Eta Kappa Nu (IEEE-HKN) CMU Chapter.

Gauri Joshi (S'12-M'17) received her B.Tech and M.Tech in Electrical Engineering from the Indian Institute of Technology (IIT) Bombay in 2010. She completed her Ph.D. from MIT EECS in June 2016. After her Ph.D. she worked as a Research Staff Member at IBM T. J. Watson Research Center. Since Fall 2017, Gauri Joshi is an assistant professor in the ECE department at Carnegie Mellon University. Her research spans distributed machine learning, large-scale parallel computing, and information theory. Her awards and honors include the NSF CAREER Award (2021), ACM Sigmetrics Best Paper Award (2020), NSF CRII Award (2018), Best Thesis Prize in Computer science at MIT (2012), and the Institute Gold Medal of IIT Bombay (2010).

Osman Yağan (S'07–M'12–SM'17) received the B.S. degree in Electrical and Electronics Engineer- ing from Middle East Technical University, Ankara (Turkey) in 2007, and the Ph.D. degree in Electrical and Computer Engineering from University of Maryland, College Park, MD in 2011. In August 2013, he joined the faculty of the Department of Electrical and Computer Engineering at Carnegie Mellon University, where he is currently an Associate Research Professor. Dr. Yagan's research is on modeling, analysis, and performance optimization of computing systems, and uses tools from applied probability, data science, machine learning, and network science. Specific

topics include wireless communications, security, random graphs, social and information networks, and cyber-physical systems. He is a recipient of the CIT Dean's Early Career Fellowship.