Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection Lee Kong Chian School Of Business

Lee Kong Chian School of Business

11-2017

Combinatorial multi-armed bandit problem with probabilistically triggered arms: A case with bounded regret

Omer SARITAC
Singapore Management University, omersaritac@smu.edu.sg

Cem TEKIN

Follow this and additional works at: https://ink.library.smu.edu.sg/lkcsb_research

Part of the Operations and Supply Chain Management Commons

Citation

SARITAC, Omer and TEKIN, Cem. Combinatorial multi-armed bandit problem with probabilistically triggered arms: A case with bounded regret. (2017). *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP): Montreal, November 14-16: Proceedings*. 111-115.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/7605

This Conference Proceeding Article is brought to you for free and open access by the Lee Kong Chian School of Business at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection Lee Kong Chian School Of Business by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Noname manuscript No.
(will be inserted by the editor)

Combinatorial Multi-armed Bandit with Probabilistically Triggered Arms: A Case with Bounded Regret*

A. Ömer Sarıtaç · Cem Tekin

the date of receipt and acceptance should be inserted later

arXiv:1707.07443v1 [cs.LG] 24 Jul 2017

Abstract In this paper, we study the combinatorial multi-armed bandit problem (CMAB) with probabilistically triggered arms (PTAs). Under the assumption that the arm triggering probabilities (ATPs) are positive for all arms, we prove that a class of upper confidence bound (UCB) policies, named Combinatorial UCB with exploration rate κ (CUCB-κ), and Combinatorial Thompson Sampling (CTS), which estimates the expected states of the arms via Thompson sampling, achieve bounded regret. In addition, we prove that CUCB-0 and CTS incur $O(\sqrt{T})$ gap-independent regret. These results improve the results in previous works, which show $O(\log T)$ gap-dependent and $O(\sqrt{T \log T})$ gap-independent regrets, respectively, under no assumptions on the ATPs. Then, we numerically evaluate the performance of CUCB-κ and CTS in a real-world movie recommendation problem, where the actions correspond to recommending a set of movies, the arms correspond to the edges between the movies and the users, and the goal is to maximize the total number of users that are attracted by at least one movie. Our numerical results complement our theoretical findings on bounded regret. Apart from this problem, our results also directly apply to the online influence maximization (OIM) problem studied in numerous prior works.

Keywords Combinatorial Multi-armed Bandit, Probabilistically Triggered Arms, Bounded Regret, Movie Recommendation, Online Influence Maximization.

1 Introduction

Multi-armed bandit (MAB) problem is a canonical example of problems that involve sequential decision making under uncertainty that has been extensively studied in the

* A preliminary version of this work will appear in IEEE GlobalSIP 2017.

A. Ömer Sarıtaç
Department of Industrial Engineering, Bilkent University, Ankara, Turkey, E-mail: omersaritac@gmail.com

Cem Tekin
Department of Electrical and Electronics Enginnering, Bilkent University, Ankara, Turkey, E-mail: cemtekin@ee.bilkent.edu.tr

past (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al, 2002; Bubeck and Cesa-Bianchi, 2012). This problem proceeds over a sequence of epochs, where the learner selects an arm in each epoch, and receives a reward that depends on the selected arm. The learner aims to maximize its cumulative reward in the long run, by estimating the arm rewards using the previous reward observations. Due to the fact that only the reward of the selected arm is revealed to the learner, in order to maximize its cumulative reward the learner needs to trade-off exploration and exploitation. In short, exploring a new arm may result in short term loss (due to not selecting the estimated best arm) but long term gain (due to discovering superior arms), while exploiting the estimated best arm may result in short term gain but long term loss (due to failing to detect superior arms).

Although theoretically appealing, the classical MAB problem described above is not appropriate for real-world applications where multiple arms are chosen in each epoch, and the resulting reward is a non-linear function of the chosen arms. These applications include wireless networking, online advertising and recommendation, and viral marketing, which are studied under the combinatorial multi-armed bandit (CMAB) formalism (Anantharam et al, 1987; Gai et al, 2012; Chen et al, 2013; Gopalan et al, 2014; Kveton et al, 2014, 2015a,b). In CMAB, the set of arms chosen by the learner at each epoch is referred to as the action. At the end of each epoch, the learner observes both the reward of the chosen action and the states of the chosen arms. This problem is significantly more difficult from the classical MAB problem due to the fact that the size of the action set is combinatorial in the number of arms.

An interesting extension to the CMAB is CMAB with PTAs (Chen et al, 2016b), where the actions chosen by the learner may trigger arms probabilistically. In this work the authors propose the combinatorial UCB (CUCB) algorithm and prove a $O(\log T)$ gap-dependent regret bound for CUCB. Later, this model is extended in Wang and Chen (2017), where the authors provide tighter regret bounds by getting rid of a problem parameter p^*, which denotes the minimum positive probability that an arm gets triggered by an action. This is achieved by introducing a new smoothness condition on the expected reward function. Based on this, the authors prove $O(\log T)$ gap-dependent and $\tilde{O}(\sqrt{T})$ gap-independent regret bounds.

In this paper, we consider an instance of CMAB with PTAs, where all of the ATPs are positive. For this problem we propose two different learning algorithms: Combinatorial UCB with exploration rate κ (CUCB-κ) and Combinatorial Thompson Sampling (CTS). The first one uses a UCB-based index to form optimistic estimates of the expected states of the arms, while the latter one samples the expected states of the arms from a posterior distribution formed using the past state observations. Then, we prove that both CUCB-κ and CTS achieve $O(1)$ gap-dependent regret for any $\kappa \geq 0$, and both CUCB-0 and CTS achieve $O(\sqrt{T})$ gap-independent regret. Here, CUCB-0 corresponds to the greedy algorithm which always exploits the best action calculated based on the sample mean estimates of the arm states. Although not very common, bounded regret appears in various MAB problems, including some instances of parameterized MAB problems (Mersereau et al, 2009; Atan et al, 2015). However, these works do not conflict with the asymptotic $O(\log T)$ lower bound for the classical MAB problem (Lai and Robbins, 1985), because in these works the reward from an arm provides information on the rewards from the other arms. We argue that bounded regret is also intuitive for our problem, because when the arms are probabilistically triggered, it is possible to observe the rewards of arms that never get selected.

The contributions of this paper are summarized as follows:

- We propose a variant of CMAB with PTAs, where all ATPs are strictly positive.
- We propose a UCB-based algorithm (CUCB-κ) and a Thompson sampling based algorithm (CTS) for CMAB with PTAs.
- We prove that the gap-dependent regrets for CUCB-κ and CTS are bounded for CMAB with PTAs, when ATPs are positive. This improves the previous $O(\log T)$ regret bound in the prior work (Chen et al, 2016b) that holds under a more general setting.
- We also prove that the gap-independent regret for CUCB-0 and CTS is $O(\sqrt{T})$ for CMAB with PTAs, when ATPs are positive. This also improves the previous $O(\sqrt{T \log T})$ regret bound in the prior work (Chen et al, 2016b) that holds under a more general setting.
- We evaluate the performance of CUCB-κ and CTS in a movie recommendation problem with PTAs defined over a bipartite graph, and illustrate how the regret is affected by the size of the action, p^*, and κ.

The rest of the paper is organized as follows. Related work is given in Section 2. Problem description is given in Section 3. Two learning algorithms are proposed in Section 4, and their regrets are analyzed in Section 5. The movie recommendation application of the CMAB with PTAs is studied in Section 6. Concluding remarks are provided in Section 7. All proofs are given in the appendix.

2 Related Work

Two key techniques are used for learning in MAB problems: UCB-based index policies (Lai and Robbins, 1985) and Thompson (posterior) sampling (Thompson, 1933). Lai and Robbins (1985) introduces UCB-based index policies for the MAB problem and proves a tight logarithmic bound on the asymptotic regret, which establishes asymptotic optimality of the proposed set of policies. Later on, Agrawal (1995) revealed that sample-mean based index policies can also achieve $O(\log T)$ regret, and Auer et al (2002) showed that $O(\log T)$ regret is achievable not only asymptotically but also uniformly over time by a very simple sample-mean based index policy. Briefly, a UCB-based index policy constructs an optimistic estimate of the expected reward of each arm by using only the reward observations gathered from that arm but not the other arms, and then, selects the arm with the highest index. Therefore, the arm selection of a UCB-based index policy is deterministic given the entire history.

Although regret bounds for UCB-based index policies exist for more than several decades, no significant progress is done for Thompson sampling until Agrawal and Goyal (2012) and Kaufmann et al (2012), which show the regret-optimality of Thompson sampling for the classical MAB problem. These efforts to prove regret bounds for Thompson sampling are motivated by works such as (Scott, 2010; Granmo, 2010; Graepel et al, 2010), which demonstrate the empirical efficiency of Thompson sampling. Unlike UCB-based index policies, Thompson sampling selects an arm by drawing samples from the posterior distribution of arm rewards. Thus, the arm selection of Thompson sampling is random given the entire history. In summary, the performance of UCB-based policies and Thompson sampling is well studied in the classical MAB problem where the learner selects one arm at a time.

On the other hand, in the CMAB problem, most of the prior works consider UCB-based policies. Among these, there exist several works where the reward function is assumed to be a linear function of the outcome of the individual arms (Auer, 2003; Dani

et al, 2008; Abbasi-Yadkori et al, 2011; Kveton et al, 2014, 2015c). In addition, Kveton et al (2015a) and Kveton et al (2015b) solve specific instances of CMAB problems with nonlinear reward functions. Moreover, similar to our work, Chen et al (2016b) and Wang and Chen (2017) consider CMAB with PTAs with a general reward function where the reward function satisfies certain bounded-smoothness and monotonicity assumptions.

There are also several works that use Thompson sampling based approaches for the CMAB problem and its variants. For instance, Gopalan et al (2014) considers a bandit problem with complex actions, where the reward of each action is a function of the rewards of the individual arms, and derives a regret bound for Thompson sampling when applied to this problem. An empirical study of Thompson sampling for the CMAB problem is carried out in Durand and Gagné (2014), where it is also used for online feature selection. In addition, recently, Thompson sampling is used in Multinomial Logit (MNL) bandit problem, which involves a combinatorial objective (Agrawal et al, 2017). To the best of our knowledge, none of these works are directly applicable to PTAs. A comparison of our regret bounds with the regret bounds derived in prior works can be found in Table 1.

A closely related problem to the CMAB problem is the influence maximization (IM) problem, which is first formulated in Kempe et al (2003) as a combinatorial optimization problem, and has been extensively studied since then. The goal in this problem is to select a seed set of nodes that maximizes the influence spread in a given network. In this problem, the seed set corresponds to the action, edges of the network correspond to arms, the influence spread corresponds to the reward, and the ATPs are determined by an influence spread process. Various works consider the online version of this problem, named the OIM problem (Lei et al, 2015; Vaswani et al, 2015; Wen et al, 2017; Sarıtaç et al, 2016). In this version, the ATPs are unknown a priori. Works such as Chen et al (2016b) and Wang and Chen (2017) solve this problem by using algorithms developed for CMAB with PTAs. Differently from these, Lei et al (2015) considers an objective function, which is given as the expected size of the union of nodes influenced in each epoch over time, Wen et al (2017) adopts the well known algorithm LinUCB for the IM problem and calls it IMLinUCB, which permits a linear generalization, making the algorithm suitable for large-scale problems. Vaswani et al (2015) introduces the node level feedback in addition to the edge level feedback used in prior works, and proposes a method that uses the node level feedback to update the estimated ATPs. In a related work, Sarıtaç et al (2016) introduce the contextual OIM problem, which is a combination of contextual bandits and OIM, and propose an algorithm that achieves sublinear regret. Importantly, the theoretical results we prove in this paper also applies to the OIM problem defined over a strongly connected graph where each node is reachable from the other nodes in the graph.

3 Problem Formulation

We adopt the notation in Wang and Chen (2017). The system operates in discrete epochs indexed by t. There are m arms, given by the set $\{1, \ldots, m\}$, whose states at each epoch are drawn from an unknown joint distribution D with support in $[0, 1]^m$. The state of arm i at epoch t is denoted by $X_i^{(t)}$, and the state vector at epoch t is denoted by $\boldsymbol{X}^{(t)} := (X_1^{(t)}, \ldots, X_m^{(t)})$.

In each epoch t, the learner selects an action S_t from the finite set of actions $\mathcal{S}$ based on its history of actions and observations. Then, a random subset of arms $\tau_t \subseteq \{1, \ldots, m\}$

Table 1: Comparison of the regret bounds in our work and the relevant literature.

	Our Work	Chen et al (2016b); Wang and Chen (2017)	Kveton et al (2014, 2015b,c); Chen et al (2016a)	Gopalan et al (2014); Kveton et al (2015a)
Gap-dependent Regret	$O(1)$	$O(\log T)$	$O(\log T)$	$O(\log T)$
Gap-independent Regret	$O(\sqrt{T})$	$O(\sqrt{T \log T})$	$O(\sqrt{T \log T})$	No Bound
Strictly positive ATPs	Yes	No	No PTAs	No PTAs

is triggered based on S_t and $\boldsymbol{X}^{(t)}$. Here, τ_t is drawn from a multivariate distribution (also called the *probabilistic triggering function*) $D^{\text{trig}}(S_t, \boldsymbol{X}^{(t)})$ with support $[p^*, 1]^m$ for some $p^* > 0$, which is equivalent to saying that all the ATPs are positive. As we will show in the subsequent sections, this key assumption allows the learner to achieve bounded regret, without the need for explicit exploration. Then, at the end of epoch t, the learner obtains a finite, non-negative reward $R(S_t, \boldsymbol{X}^{(t)}, \tau_t)$ that depends deterministically on S_t, $\boldsymbol{X}^{(t)}$ and τ_t, and observes the states of the triggered arms, i.e., $X_i^{(t)}$, $i \in \tau_t$. The goal of the learner is to maximize its total expected reward over all epochs.

For each arm $i \in \{1, \ldots, m\}$, we let $\mu_i := \mathbb{E}_{\boldsymbol{X}^{(t)} \sim D}[X_i^{(t)}]$ denote the expected state of arm i and $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_m)$ denote the *expectation vector*. The expected reward of action S is $r_{\boldsymbol{\mu}}(S) := \mathbb{E}[R(S, \boldsymbol{X}, \tau)]$ where the expectation is taken over $\boldsymbol{X} \sim D$ and $\tau \sim D^{\text{trig}}(S, \boldsymbol{X})$. We call $r_{\boldsymbol{\mu}}(\cdot)$ the *expected reward function*. Let S^* denote an optimal action such that $S^* \in \arg\max_{S \in \mathcal{S}} r_{\boldsymbol{\mu}}(S)$. The expected reward of the optimal action is given by $r^*_{\boldsymbol{\mu}}$.

Computing the optimal action even when the expected states and the probabilistic triggering function are known is often an NP-hard problem for which (α, β)-approximation algorithms exist (Vazirani, 2001). Due to this, we compare the performance of the learner with respect to an (α, β)-approximation algorithm $\mathcal{O}$, which takes $\boldsymbol{\mu}$ as input and outputs an action $S^{\mathcal{O}}$ such that $\Pr(r_\mu(S^{\mathcal{O}}) \geq \alpha r^*_{\boldsymbol{\mu}}) \geq \beta$. Here, α denotes the approximation ratio and β denotes the minimum success probability. Based on this, the (α, β)-approximation regret (simply referred to as the regret) of the learner that uses a learning algorithm π to select actions by epoch T is defined as follows:

$$\text{Reg}^{\pi}_{\boldsymbol{\mu},\alpha,\beta}(T) := T\alpha\beta r^*_{\boldsymbol{\mu}} - \mathbb{E}\left[\sum_{i=1}^{T} r_{\boldsymbol{\mu}}(S_t)\right]. \tag{1}$$

For the purpose of regret analysis, as in Chen et al (2016b), we impose two mild assumptions on the expected reward function. The first assumption states that the expected reward function is smooth and bounded.

Assumption 1 (Chen et al (2016b))**.** *$\exists f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ such that f is continuous, strictly increasing, and $f(0) = 0$, where f is called the bounded smoothness*

function. For any two expectation vectors, $\boldsymbol{\mu}$ *and* $\boldsymbol{\mu}'$*, and for any* $\Delta > 0$*, we have* $|r_{\boldsymbol{\mu}}(S) - r_{\boldsymbol{\mu}'}(S)| \leq f(\Delta)$*, if* $\max_{i\in\{1,\ldots,m\}} |\mu_i - \mu_i'| \leq \Delta$*,* $\forall S \in \mathcal{S}$*.*

The second assumption states that the expected reward is monotone under $\boldsymbol{\mu}$.

Assumption 2 (Chen et al (2016b))**.** *If for all arms* $i \in \{1,\ldots,m\}$*,* $\mu_i \leq \mu_i'$*, then we have* $r_{\boldsymbol{\mu}}(S) \leq r_{\boldsymbol{\mu}'}(S)$*,* $\forall S \in \mathcal{S}$*.*

4 Learning Algorithms

In this section we propose a UCB-based learning algorithm called Combinatorial UCB with exploration rate κ (CUCB-κ) and a Thompson sampling based learning algorithm called Combinatorial Thompson Sampling (CTS) for the CMAB problem with PTAs.

4.1 CUCB-κ

Algorithm 1 Combinatorial UCB-κ (CUCB-κ)

Input: Set of actions $\mathcal{S}$, $\kappa > 0$
Initialize counters: For each arm $i \in \{1,\ldots,m\}$, set $T_i = 0$, which is the number of times arm i is observed. $t = 1$
Initialize estimates: Set $\overline{\mu}_i^{\kappa} = 1$ and $\hat{\mu}_i = 1$, $\forall i \in \{1,\ldots,m\}$, which are the UCB and sample mean estimates for μ_i, respectively
while $t \geq 1$ **do**
  Call the (α,β)-approximation algorithm with $\overline{\boldsymbol{\mu}}^{\kappa}$ as input to get S_t
  Select action S_t, observe $X_i^{(t)}$'s for $i \in \tau_t$ and collect the reward R
  for $i \in \tau_t$ **do**
    $T_i = T_i + 1$
    $\hat{\mu}_i = \hat{\mu}_i + \frac{X_i^{(t)} - \hat{\mu}_i}{T_i}$
  end for
  for $i \in \{1,\ldots,m\}$ **do**
    $\overline{\mu}_i^{\kappa} = \min\left\{\hat{\mu}_i + \kappa\sqrt{\frac{3\ln t}{2T_i}}, 1\right\}$
  end for
  $t = t + 1$
end while

The pseudocode of CUCB-κ is given in Algorithm 1. CUCB-κ is almost the same as CUCB algorithm given in Chen et al (2016b), with the exception that the inflation term (adjustment term) is multiplied by a scaling factor $\kappa \geq 0$. CUCB-κ keeps a counter T_i, which tracks the number of times each arm i is played as well as the sample mean and the UCB estimate of its expected states, denoted by $\hat{\mu}_i$ and $\overline{\mu}_i^{\kappa}$, respectively. Let $\hat{\boldsymbol{\mu}} := \{\hat{\mu}_1,\ldots,\hat{\mu}_m\}$ and $\overline{\boldsymbol{\mu}}^{\kappa} := \{\overline{\mu}_1^{\kappa},\ldots,\overline{\mu}_m^{\kappa}\}$ denote the estimated expectation vector and the UCB for the expectation vector, respectively. We will use superscript t when explicitly referring to the counters and estimates that CUCB-κ uses at epoch t. For instance, we have $T_i^t = \sum_{j=1}^{t-1} 1_{\{i\in\tau_j\}}$, $\hat{\mu}_i^t = \frac{1}{T_i^t}\sum_{j=1}^{t-1} X_i^{(j)} 1_{\{i\in\tau_j\}}$ and

$\overline{\mu}_i^{\kappa,t} = \hat{\mu}_i^t + \min\{\kappa\sqrt{\frac{3\log t}{2T_i^t}}, 1\}$. Initially, CUCB-$\kappa$ sets T_i^1 as 0 and $\overline{\mu}_i^{\kappa,1}$ as 1 for all arms. Then, in each epoch $t \geq 1$, it calls an (α, β)-approximation algorithm, which takes as input $\overline{\boldsymbol{\mu}}^{\kappa,t}$ and chooses an action S_t. The action S_t depends on the randomness of the approximation algorithm itself in addition to $\overline{\boldsymbol{\mu}}^{\kappa,t}$. After playing the action S_t, the states of the arms in $i \in \tau_t$, i.e., $\{X_i^{(t)}\}_{i\in\tau_t}$, are revealed, and a reward R that depends on S_t, $\boldsymbol{X}^{(t)}$ and τ_t is collected by the learner. Then, CUCB-κ updates its estimates $\hat{\boldsymbol{\mu}}^{t+1}$ and $\overline{\boldsymbol{\mu}}^{\kappa,t+1}$ for the next epoch based on τ_t and $\{X_i^{(t)}\}_{i\in\tau_t}$.

When $\kappa = 0$, the inflation term in CUCB-κ vanishes. In this case, the algorithm always selects an action that is produced by an (α, β)-approximation algorithm that takes as input the estimated expectation vector. Hence, CUCB-κ becomes a greedy algorithm that always exploits based on the current values of the estimated parameters. Although such an algorithm will incur high regret in the classical MAB problem, we will show that CUCB-0 performs surprisingly well in our problem due to the fact that the ATPs are all positive.

4.2 CTS

Algorithm 2 CTS

Input: Set of actions $\mathcal{S}$
Initialize counters: For each arm $i \in \{1, \ldots, m\}$, set $s_i = 0$ and $f_i = 0$, which are the success and failure counts of arm i. $t = 1$
while $t \geq 1$ **do**
  For each arm $i \in \{1, 2, \ldots, m\}$, sample ν_i from the Beta(s_i+1, f_i+1) distribution.
  Call the (α, β)-approximation algorithm with $\boldsymbol{\nu}$ as input to get S_t.
  Select action S_t, observe $X_i^{(t)}$'s for $i \in \tau_t$ and collect the reward R.
  for $i \in \tau_t$ **do**
    if $X_i^{(t)} = 1$ **then**
      $s_i = s_i + 1$
    else
      $f_i = f_i + 1$
    end if
  end for
  $t = t + 1$
end while

The pseudocode of CTS is given in Algorithm 2. For the simplicity of exposition, for CTS, in addition to the assumptions given in Section 3, we also assume that $X_i^{(t)} \in \{0, 1\}$ for all $i \in \{1, \ldots, m\}$, i.e., the states of the arms are Bernoulli random variables. Note that CTS can easily be generalized to the case when $X_i^{(t)} \in [0, 1]$ for all $i \in \{1, \ldots, m\}$, by performing a Bernoulli trial for any arm $i \in \tau_t$ with success probability $X_i^{(t)}$, in way similar to the extension described in Agrawal and Goyal (2012). Also note that Bernoulli arm states are very common, and appear in the OIM problem and the movie recommendation example that we discuss in this paper.

For each arm i, CTS keeps two counters s_i and f_i, which count the number of times the state of arm i is observed as 1 (success) and 0 (failure), respectively. We denote by $\boldsymbol{\nu} := \{\nu_1, \ldots, \nu_m\}$, the estimated expectation vector where ν_i is drawn from $\text{Beta}(s_i + 1, f_i + 1)$ in each epoch. Similar to CUCB-κ, we use superscripts when explicitly referring to the counters and estimates that CTS uses in epoch t. For instance, ν_i^t denotes a sample drawn from $\text{Beta}(s_i^t + 1, f_i^t + 1)$, where s_i^t and f_i^t are the values of the counters s_i and f_i in epoch t.

Initially, CTS sets $s_i = 0$ and $f_i = 0$. In each epoch $t \geq 1$, it takes a sample ν_i from the distribution $\text{Beta}(s_i + 1, f_i + 1)$ for each arm $i \in \{1, \ldots, m\}$, which is used as an estimate for μ_i. Then, it calls an (α, β)-approximation algorithm, which takes as input the estimates $\boldsymbol{\nu}$ for the expectation vector $\boldsymbol{\mu}$ and chooses an action S_t. The action S_t depends on the randomness of the approximation algorithm itself in addition to $\boldsymbol{\nu}$. After playing the action S_t, $\{X_i^{(t)}\}_{i \in \tau_t}$ is revealed, and the learner collects the reward R just as in CUCB-κ. Then, CTS updates its counters s_i and f_i for all arms $i \in \tau_t$. If $X_i^{(t)} = 1$, then s_i is incremented by one. Otherwise, if $X_i^{(t)} = 0$, then f_i is incremented by one. The counters of the arms that are not in τ_t remain unchanged.

5 Regret Analysis

In this section we analyze the regrets of CUCB-κ and CTS. Before delving into the details of the regret analysis, we first prove a key theorem, which shows that the event that the number of times an arm is played by the end of epoch t is less than a linear function of t for some arm has a very low probability for t sufficiently large.

Theorem 1. *For any learning algorithm, $\eta \in (0, 1)$ and for all natural numbers $t \geq t' := 4c^2/e^2$, where $c := 1/(p^*(1-\eta))^2$, we have*

$$\Pr\left(\bigcup_{i \in \{1,\ldots,m\}} \left\{T_i^{t+1} \leq \eta p^* t\right\}\right) \leq \frac{m}{t^2}.$$

Theorem 1 is the crux of achieving the theoretical results in this paper since it guarantees that any algorithm obtains sufficiently many observations from each arm, including algorithms that do not explicitly explore any of the arms. This result is very intuitive because when all of the ATPs are positive, the learner observes the states of all arms with positive probability for any action it selects. This will allow us to prove that the estimated expectation vector converges to the true expectation vector independent of the learning algorithm that is used. This fact will be used in proving that the gap-dependent regrets of CUCB-κ and CTS are bounded.

Before continuing the regret analysis, we provide some additional notation. Let n_B denote the number of actions whose expected rewards are smaller than $\alpha r_{\boldsymbol{\mu}}^*$. These actions are called *bad actions*. We re-index the bad actions in increasing order such that $S_{B,l}$ denotes the bad action with lth smallest expected reward. The set of bad actions is denoted by $\mathcal{S}_B := \{S_{B,1}, S_{B,2}, \ldots, S_{B,n_B}\}$. Let $\nabla_l := \alpha r_{\boldsymbol{\mu}}^* - r_{\boldsymbol{\mu}}(S_{B,l})$ for each $l \in \{1, \ldots, n_B\}$ and $\nabla_{n_B+1} = 0$. Accordingly, we let $\nabla_{\max} := \nabla_1$, $\nabla_{\min} := \nabla_{n_B}$. We also let $\text{gap}(S_t) := \alpha r_{\boldsymbol{\mu}}^* - r_{\boldsymbol{\mu}}(S_t)$.

5.1 Regret Analysis for CUCB-κ

First, we show that, given any constant $\delta > 0$, the probability that

$$\Delta_t^\kappa := \max_{i \in \{1,\ldots,m\}} |\mu_i - \overline{\mu}_i^{\kappa,t}| < \delta$$

is high, when t is sufficiently large. This measures how well CUCB-κ learns the expected state of each arm by the beginning of epoch t, and is directly related to Theorem 1 as it is related to the number of times each arm is observed by epoch t.

Theorem 2. *Consider CUCB-κ, where $\kappa > 0$. For any $\delta > 0$ and $\eta \in (0,1)$, let $c := 1/(p^*(1-\eta))^2$, $c_0 := 6\kappa^2/(\delta^2 p^* \eta)$ and $t_1 := \max\{4c^2/e^2, 4c_0^2/e^2\}$. When CUCB-$\kappa$ is run, we have for all integers $t \geq t_1$*

$$\Pr(\Delta_{t+1}^\kappa \geq \delta) \leq \frac{2m}{t^2(1-e^{-\delta^2/2})} + 2me^{-\delta^2 \eta p^* t/2} + \frac{m}{t^2}.$$

Consider CUCB-0. For any $\delta > 0$ and $\eta \in (0,1)$, let $c := 1/(p^*(1-\eta))^2$ and $t' := 4c^2/e^2$. When CUCB-0 is run, we have for all integers $t \geq t'$

$$\Pr(\Delta_{t+1}^0 \geq \delta) \leq \frac{2m}{t^2(1-e^{-2\delta^2})} + 2me^{-2\delta^2 \eta p^* t}.$$

The upper bound for CUCB-κ, $\kappa > 0$ is looser than the upper bound for CUCB-0 given in Theorem 2, because of the fact that $t_1 \geq t'$ and additional m/t^2 term that appears in the upper bound for CUCB-κ, $\kappa > 0$. These terms appear as an artifact of the presence of the additional inflation term $\kappa\sqrt{\frac{3\ln t}{2T_i}}$ that appears in the UCB for the expectation vector. While this observation about the upper bound is not sufficient to conclude that CUCB-κ, $\kappa > 0$ is worse than CUCB-0 in the setting that we consider, our empirical finding in Section 6 shows that CUCB-0 incurs smaller regret than CUCB-κ, $\kappa > 0$ for the movie recommendation application that we consider.

The next theorem shows that the regret of CUCB-κ is bounded for any $T > 0$.

Theorem 3. *The regret of CUCB-κ, $\kappa > 0$ is bounded, i.e., $\forall T \geq 1$*

$$Reg_{\boldsymbol{\mu},\alpha,\beta}^{CUCB\text{-}\kappa}(T) \leq \nabla_{\max} \inf_{\eta \in (0,1)} \left(\lceil t_1 \rceil + \frac{m\pi^2}{3}\left(\frac{2}{\delta^2} + \frac{3}{2}\right) + 2m\left(1 + \frac{2}{\delta^2 \eta p^*}\right)\right) \quad (2)$$

where $\delta := f^{-1}(\nabla_{\min}/2)$, $t_1 := \max\{4c^2/e^2, 4c_0^2/e^2\}$, $c := 1/(p^*(1-\eta))^2$ and $c_0 := 6\kappa^2/(\delta^2 \eta p^*)$.

The regret of CUCB-0 is bounded, i.e., $\forall T \geq 1$

$$Reg_{\boldsymbol{\mu},\alpha,\beta}^{CUCB\text{-}0}(T) \leq \nabla_{\max} \inf_{\eta \in (0,1)} \left(\lceil t' \rceil + \frac{m\pi^2}{3}\left(1 + \frac{1}{2\delta^2}\right) + 2m\left(1 + \frac{1}{2\delta^2 \eta p^*}\right)\right) \quad (3)$$

where $\delta := f^{-1}(\nabla_{\min}/2)$, $t' := 4c^2/e^2$ and $c := 1/(p^*(1-\eta))^2$.

This result is different from the prior results (Chen et al, 2016b; Wang and Chen, 2017; Kveton et al, 2015b) where $O(\log T)$ gap-dependent regret upper bounds are proven for the CMAB problem. The main difference of our problem from these works is that we assume the minimum ATP to be positive. This allows us to prove the result in Theorem 1, by ensuring that each arm is triggered sufficiently many times independent of the exploration strategy used by the learner.

When $\nabla_{\min}$ becomes too small in Theorem 3, the regret approaches to infinity because when $\nabla_{\min}$ is too small, δ will also become very small. For a large class of problems where the bounded smoothess function is $f(x) = \gamma x^w$ where $w \in (0, 1]$, $\gamma > 0$, and $\kappa = 1$, the worst case regret is shown to be $O(T^{1-\omega/2}(\ln T)^{\omega/2})$ in Chen et al (2016b) for a more general setting, hence, this regret bound also holds for our problem. We show in Theorem 4 that for $\kappa = 0$ (when CUCB-κ becomes the greedy policy), the worst-case regret is bounded by $O(T^{1-w/2})$. Note that for the OIM problem and the recommendation problem we consider in the experiments, we have $\omega = 1$, and hence, this bound becomes $O(\sqrt{T})$. To prove this, we investigate the behavior of $\text{gap}(S_t) = \alpha r^*_{\boldsymbol{\mu}} - r_{\boldsymbol{\mu}}(S_t)$ based on the change in Δ^0_t. For this, we use Theorem 2 to bound the expected value of Δ^0_t, which allows us to bound the gap-independent regret.

Theorem 4. *When the bounded-smoothness function in Assumption 1 is $f(x) = \gamma x^w$ where $\gamma > 0$ and $\omega \in (0, 1]$, the gap-independent regret bound for CUCB-0 is*

$$Reg^{CUCB\text{-}0}_{\boldsymbol{\mu},\alpha,\beta}(T) \leq \inf_{\eta\in(0,1)} \left(\lceil t' \rceil \nabla_{max} + \gamma(2m)^\omega \left[2^\omega \left(\frac{\pi}{2\eta p^*} \right)^{\omega/2} + 3^\omega \right] \frac{T^{1-w/2}}{1-w/2} \right)$$

where $t' := 4c^2/e^2$ and $c := 1/(p^*(1-\eta))^2$. Hence, the gap-independent regret of CUCB-0 is $O(T^{1-\frac{w}{2}})$.

As a remark, note that the gap-independent regret bound holds for all problem instances where the minimum ATP is at least p^*. Essentially, gap-independent means that the regret bound does not depend on $\nabla_{\min}$, and hence δ. Also, $\nabla_{\max}$ can be bounded by the maximum reward, since the reward is assumed to be finite.

5.2 Regret Analysis for CTS

The analysis of the regret of CTS is similar to the regret analysis for CUCB-κ. We first show that the probability that $\Delta^{\boldsymbol{\nu}}_t := \max_i |\nu^t_i - \mu_i|$ is greater than some constant $\delta > 0$ becomes smaller as t increases.

Theorem 5. *When CTS is run, for any $\delta > 0$ and $\eta \in (0, 1)$, we have for all integers $t \geq t' := 4c^2/e^2$*

$$\Pr(\Delta^{\boldsymbol{\nu}}_{t+1} \geq \delta) \leq (3 + e^{2\delta})m \left(\frac{1}{t^2(1 - e^{-\delta^2/2})} + e^{-\delta^2 \eta p^* t/2} \right)$$

where $c = 1/(p^*(1-\eta))^2$.

The next theorem proves that the regret of CTS is bounded for any $T > 0$.

Theorem 6. *The regret of CTS is bounded, i.e., $\forall T \geq 1$*

$$\begin{aligned} &Reg^{CTS}_{\boldsymbol{\mu},\alpha,\beta}(T) \\ &\leq \nabla_{\max} \inf_{\eta\in(0,1)} \left(\lceil t' \rceil + \frac{(3 + e^{2\delta})m\pi^2}{6} \left(1 + \frac{2}{\delta^2} \right) + (3 + e^{2\delta})m \left(1 + \frac{2}{\delta^2 \eta p^*} \right) \right). \end{aligned}$$

where $\delta := f^{-1}(\nabla_{\min}/2)$, $t' := 4c^2/e^2$ and $c := 1/(p^*(1-\eta))^2$.

The next theorem gives the gap-independent regret bound for CTS.

Theorem 7. *When the bounded-smoothness function in Assumption 1 is $f(x) = \gamma x^w$ where $\gamma > 0$ and $\omega \in (0, 1]$, the gap-independent regret bound for CTS is*

$$Reg^{CTS}_{\boldsymbol{\mu},\alpha,\beta}(T) \leq \inf_{\eta \in (0,1)} \left(\lceil t' \rceil \nabla_{\max} + \gamma \Big(2m(3 + e^2) \Big)^{\omega} \left[\left(\frac{2\pi}{\eta p^*} \right)^{\omega/2} + 3^{\omega} \right] \frac{T^{1-w/2}}{1 - w/2} \right)$$

where $t' := 4c^2/e^2$ and $c := 1/(p^*(1-\eta))^2$. Hence, the gap-independent regret of CTS is $O(T^{1-\frac{w}{2}})$.

Similar to the previous case, for the OIM problem, the bound in Theorem 7 becomes $O(\sqrt{T})$.

6 Illustrative Results

In this section, we evaluate the performance of CUCB-κ and CTS on a recommendation problem. This problem has become popular among researchers with the popularization of on-demand media streaming services like Netflix. We use the *MovieLens* dataset for our experiments. The dataset contains 138k people who assigned 20M ratings to 27k movies between January 1995 and March 2015. We use the portion of the dataset that was collected between March 2014 and March 2015, which consists of 750k ratings. For our experiments, we choose 200 movies in total among the movies that were rated more than 200 times: 50 movies with the smallest ratings, 50 movies with the highest ratings, and 100 movies randomly.

6.1 Definition of the Recommendation Problem

The problem consists of a weighted bipartite graph $G = (L, R, E, p)$ where L denotes the set of movies, R denotes the set of users,[1] E denotes the set of edges between the users, and $p = \{p_{i,j}\}_{(i,j)\in E}$, where $p_{i,j}$ is the weight of edge (i, j), which corresponds to the probability that movie i influences (attracts) user j. The goal of the learner is to find a set $S \subseteq L$ of size k that maximizes the expected number of attracted nodes in R. This problem is an instance of the probabilistic maximum coverage problem (Chen et al, 2016b). Our problem extends this problem by allowing the nodes in S to trigger any $(i, j) \in E$ probabilistically. For instance, this can happen if the users also interact with each other in a social network, where the recommendation made to a user in the network may influence other users into watching the recommended movie via the word of mouth effect. Moreover, both the triggering and influence probabilities are initially unknown. We let $p_S^{i,j}$ denote the probability that action S triggers edge $(i, j) \in E$. The expected reward is defined as the expected total number of users that are attracted by at least one movie, and is given as $r_G(S) = \sum_{j\in R}(1 - \prod_{(i,j)\in E}(1 - p_S^{i,j} p_{i,j}))$. We assume that $p_S^{i,j} = 1$ for the outgoing edges of nodes $i \in S$. This assumption merely says that user j will watch movie i with probability $p_{i,j}$ when the movie is recommended to the user by the system. For the nodes $i \notin S$, $p_S^{i,j} < 1$. For these nodes, $p_S^{i,j}$ denotes the probability that user i gets to know about movie j by the word of mouth affect, without the recommender system showing the movie to the user. For simulations, we

[1] Each user corresponds to a pool of individuals with same type of preferences over genres.

set $p_S^{i,j} = p^*$, and evaluate the effect of different values of p^* on the performance of CUCB-κ and CTS.

The above problem can be viewed as an instance of CMAB with PTAs. Each edge $(i,j) \in E$ is an arm and the state of each arm is a Bernoulli random variable with success probability $p_{i,j}$. For this problem, Assumption 1 is satisfied with the bounded-smoothness function $f(x) = |E|x$. The monotonicity assumption is also satisfied for this problem since increasing the $p_{i,j}$'s will definitely increase the expected reward. In addition, the reward function is submodular, and hence, it can be shown that using the greedy algorithm in Nemhauser et al (1978), we can achieve $(1-1/e)$-approximation to the optimal reward. Hence, the greedy algorithm can be used as a $(1-1/e, 1)$-approximation algorithm.

6.2 Calculation of the Influence Probabilities

The MovieLens dataset contains the following attributes for each user: UserId, MovieId, Rating, TimeStamp, Title, and the Genre. Hence, we have the rating each user assigned to a movie with a particular genre and title. The dataset contains 20 genres. For each user $j \in R$ we first calculate the user preference vector $\mathbf{u_j}$, which is a unit vector, where each element of the vector corresponds to a coefficient representing how much the user likes a particular genre. We assume that the genre distribution of the movies that the users rated represents their genre preferences. Note that a movie can have multiple genres. We also create a 20 dimensional vector $\mathbf{g_i}$ for each movie i, and let $\mathbf{g_i}_k = 1$ if a movie belongs to genre k and 0 otherwise. Using this vector, we calculate the genre preference vector $\mathbf{u_j} = \frac{\sum_{i \in L} \mathbf{g_i} + \epsilon_{i,j}}{||\sum_{i \in L} \mathbf{g_i} + \epsilon_{i,j}||}$ for each user $j \in R$, where $\epsilon_{i,j} \sim \text{Half-Normal}(\sigma = 0.05)$. The role of $\epsilon_{i,j}$ here is to account for the fact that the user may possibly explore new genres. Similarly, for each movie $i \in L$, we calculate the unit movie genre vector $\mathbf{m_i}$ as $\mathbf{g_i}/||\mathbf{g_i}||$. Using these, the influence probabilities are calculated as $p_{i,j} = sc \times \frac{<\mathbf{m_i}, \mathbf{u_j}> r_i}{\max r_i}$, $(i,j) \in E$, where r_i is the average rating given by all users to the movie i and sc is a scale factor in $(0,1]$. This way, we took into account the quality in addition to the type (genre) of the movies in determining the influence probabilities.

6.3 Results

All of the presented results are for $p^* = 0.05$, $k = 16$, $sc = 0.2$, and $\kappa = 0$ unless otherwise stated. In addition, to be able to make plausible comparisons between settings with different parameters, we consider a scaled version of the regret, where the regret is divided by the $\alpha\beta$ fraction of the optimal reward. $\alpha\beta$ fraction of the optimal reward is calculated by running the (α, β)-approximation algorithm, which is the greedy algorithm from Nemhauser et al (1978), by giving the true influence probabilities as input. We observe from Fig. 1 that the regret is bounded for different values of k and p^* for both of the algorithms and that the regret is bounded for different values of κ for CUCB-κ. It is observed that both CUCB-κ and CTS incurs almost no regret after the first 300 epochs. Moreover, for both of the algorithms, as p^* or k increases, the regret becomes smaller. On the other hand, for CUCB-κ, the regret becomes larger as κ increases, which shows that exploration hurts rather than it helps in this setting. Another observation is that although the regret of CTS and CUCB-0 increases as p^* or k decreases, CTS is affected more than CUCB-0 by the changes in p^* or k.

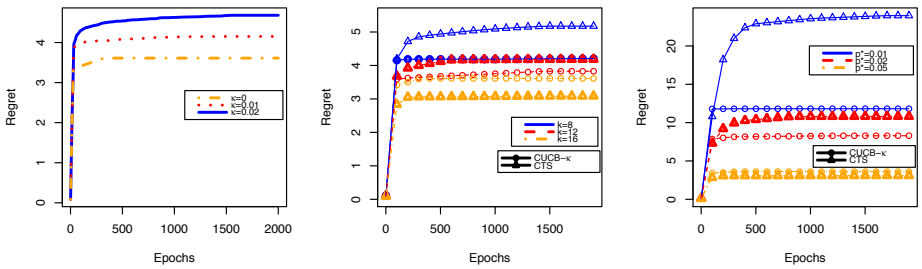

Fig. 1: Regrets of CUCB-κ and CTS for different parameter values. Left figure: CUCB-κ for $\kappa = 0, 0.01, 0.02$. Middle figure: CUCB-κ for $\kappa = 0$ (circle marker) and CTS (triangle marker). Right figure: CUCB-κ for $\kappa = 0$ (circle marker) and CTS (triangle marker).

7 Conclusion

In this paper we consider the CMAB problem with positive ATPs, and prove that CUCB-κ and CTS achieve bounded gap-dependent regret for any number of epochs T. In addition, we prove that CTS and CUCB-0 incur at most $O(\sqrt{T})$ gap-independent regret. We also show numerically that CUCB-κ and CTS achieve bounded regret in a real-world movie recommendation problem. These results suggest that exploration strategies may not be necessary for learning algorithms that work in CMAB with PTAs where ATPs are positive.

8 Appendix

8.1 Preliminaries

First, we define the instantaneous (α, β)-approximation regret, which will be used throughout the analysis.

Definition 1. *The instantaneous (α, β)-approximation regret of algorithm π at epoch t is given as*

$$IR^{\pi}_{\boldsymbol{\mu},\alpha,\beta}(t) = \alpha\beta r^*_{\boldsymbol{\mu}} - r_{\boldsymbol{\mu}}(S_t).$$

Next, we define the *nice event* at epoch t, denoted by $\mathcal{N}_t$.

Definition 2. *$\mathcal{N}_t$ is the event when the (α, β)-approximation algorithm yields a reward greater than or equal to $\alpha\beta r^*_{\boldsymbol{\mu}}$ in expectation, i.e., $\mathbb{E}[r_{\boldsymbol{\mu}}(S_t)|\mathcal{N}_t] \geq \alpha\beta r^*_{\boldsymbol{\mu}}$.*

Let $\hat{\mu}_i^{(\alpha,\beta),t}$ be the estimate of the expected state of arm i given to the (α, β)-approximation algorithm in epoch t and $\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t} := (\hat{\mu}_1^{(\alpha,\beta),t}, \ldots, \hat{\mu}_m^{(\alpha,\beta),t})$. For instance, for CUCB-κ and CTS, $\hat{\mu}_i^{(\alpha,\beta),t} = \bar{\mu}_i^{\kappa,t}$ and $\hat{\mu}_i^{(\alpha,\beta),t} = \nu_i^{\kappa,t}$, respectively. Let $\Delta_t^{(\alpha,\beta)} := \max_{i\in\{1,\ldots,m\}} |\mu_i - \hat{\mu}_i^{(\alpha,\beta),t}|$. The next lemma provides a connection between $\Delta_t^{(\alpha,\beta)}$ and the performance of the (α, β)-approximation algorithm.

Lemma 1 *Given $\Delta_t^{(\alpha,\beta)} < f^{-1}(\nabla_{\min}/2)$, an (α, β)-approximation algorithm will select an action for which $gap(S_t) \leq 0$, with probability at least β.*

Proof. Let $\hat{S}_t^* := \arg\max_{S\in\mathcal{S}} r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(S)$ be the optimal action given the estimated expectation vector $\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}$. Using Assumption 1 and the fact that $r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(\hat{S}_t^*) \geq r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(S^*)$, we get

$$\begin{aligned} r_{\boldsymbol{\mu}}(S_t) &\geq r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(S_t) - f(\Delta_t^{(\alpha,\beta)}) \\ &\geq \alpha r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(\hat{S}_t^*) - f(\Delta_t^{(\alpha,\beta)}) \\ &\geq \alpha r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(S^*) - f(\Delta_t^{(\alpha,\beta)}) \\ &\geq \alpha r_{\boldsymbol{\mu}}^* - 2f(\Delta_t^{(\alpha,\beta)}) \end{aligned} \tag{4}$$

with probability at least β. Next, we show that (4) and $S_t \in \{S_{B,1},\ldots,S_{B,n_B}\}$ cannot hold at the same time. $S_t \in \{S_{B,1},\ldots,S_{B,n_B}\}$ and $\{r_{\boldsymbol{\mu}}(S_t) \geq \alpha r_{\boldsymbol{\mu}}^* - 2f(\Delta_t^{(\alpha,\beta)})\}$ implies that

$$\alpha r_{\boldsymbol{\mu}}^* - \nabla_{\min} \geq r_{\boldsymbol{\mu}}(S_t) \geq \alpha r_{\boldsymbol{\mu}}^* - 2f(\Delta_t^{(\alpha,\beta)}).$$

The above set of inequalities cannot hold when $\Delta_t^{(\alpha,\beta)} < f^{-1}(\nabla_{\min}/2)$, since in this case we have $\nabla_{\min} > 2f(\Delta_t^{(\alpha,\beta)})$. This implies that $S_t \notin \{S_{B,1},\ldots,S_{B,n_B}\}$ with probability at least β when $\Delta_t^{(\alpha,\beta)} < f^{-1}(\nabla_{\min}/2)$. □

8.2 Proof of Theorem 1

The following lemma shows that the number of times an arm is triggered increases linearly in t with probability at least $1 - 1/t^2$.

Lemma 2 *For any learning algorithm and for any $\eta \in (0,1)$, we have*

$$\Pr(T_i^{t+1} \leq \eta p^* t) \leq \frac{1}{t^2}$$

for all integers $t \geq t' := 4c^2/e^2$, where $c = 1/(p^*(1-\eta))^2$.

Proof. In the proof, we follow a procedure similar to the proof of Theorem 3 in Akbarzadeh and Tekin (2016). Let $\mathbf{1}_{\{i\in\tau_j\}}$ be the indicator variable, which is 1 if arm i is triggered in epoch j and 0 otherwise. Recalling that $\sum_{j=1}^t \mathbf{1}_{\{i\in\tau_j\}} = T_i^{t+1}$ and using the Hoeffding's inequality, we obtain

$$\Pr(T_i^{t+1} - \mathbb{E}[T_i^{t+1}] \leq -z) \leq e^{-2z^2/t}$$

for $z > 0$. By setting $z = \sqrt{t\ln t}$, we obtain

$$\Pr(T_i^{t+1} - \mathbb{E}[T_i^{t+1}] \leq -\sqrt{t\ln t}) \leq e^{-2\ln t} = \frac{1}{t^2}.$$

Note that $\Pr(i \in \tau_t) = \mathbb{E}[\mathbf{1}_{\{i\in\tau_t\}}] \geq p^*$ for any $t \in \mathbb{N}_+$, and hence, $\mathbb{E}[T_i^{t+1}] \geq p^* t$. Following this observation, we obtain that $\{T_i^{t+1} - p^* t \leq -\sqrt{t\ln t})\} \subseteq \{T_i^{t+1} - \mathbb{E}[T_i^{t+1}] \leq -\sqrt{t\ln t})\}$. Hence, we have

$$\Pr(T_i^{t+1} - p^* t \leq -\sqrt{t\ln t}) \leq \frac{1}{t^2}.$$

We want to show that, given any $\eta \in (0,1)$, for t sufficiently large, the inequality $p^*t - \sqrt{t \ln t} \geq \eta p^* t$ holds. Let $h(t) := p^*t(1-\eta)$ and $g(t) := \sqrt{t \ln t}$. Let $t^+ \in \mathbb{R}$ be the greatest root of the equation $h(t) = g(t)$, i.e., for all $t > t^+$, $h(t) > g(t)$. To see this, we observe that $h(1) > g(1)$, and $h'(t) = p^*(1-\eta)$ and $g'(t) = \frac{1}{2}[\sqrt{\frac{\ln t}{t}} + \sqrt{\frac{1}{t \ln t}}]$. Therefore $\lim_{t\to\infty} g(t) = 0$ whereas $h'(t) = p^*(1-\eta) \in \mathbb{R}^+$ for all $t \in \mathbb{N}_+$. As a result, $\exists t'' \in \mathbb{R}$ such that $h(t) > g(t)$, $\forall t \geq t''$. This justifies that t^+ finite.

Next, we find an upper bound for t^+. Note that $h(t) = g(t)$ implies $t(p^*(1-\eta))^2 = \ln t$. Let $c = 1/(p^*(1-\eta))^2$. Thus, taking exponentials of the both sides, we want to solve the equation $e^{t/c} = t$ whose roots are given by the equation

$$t = \frac{c}{\ln e} \text{glog}\left(\frac{c}{\ln e}\right) = c\text{glog}(c).$$

Then, using the bound given in Kalman (2001), we obtain

$$t^+ \leq c^2 \left(\frac{k}{e}\right)^{\frac{k}{k-1}} \quad \text{for } k > 1.$$

By setting $k = 2$, we obtain the following upper bound on t^+:

$$t^+ \leq \frac{4c^2}{e^2}. \tag{5}$$

This implies that $\forall t \geq 4c^2/e^2 \geq t^+$, we have $h(t) \geq g(t)$, which further implies that $p^*t - \sqrt{t \ln t} \geq \eta p^* t$. Thus, for all integers $t \geq 4c^2/e^2$, we have $\{T_i^{t+1} \leq \eta p^* t\} \subseteq \{T_i^{t+1} \leq p^*t - \sqrt{t \ln t}\}$ and we obtain

$$\Pr(T_i^{t+1} \leq \eta p^* t) \leq \frac{1}{t^2}.$$

□

Using the result of Lemma 2 and the union bound over all arms, we obtain

$$\Pr\left(\bigcup_{i \in \{1,\ldots,m\}} \{T_i^{t+1} \leq \eta p^* t\}\right) \leq \frac{m}{t^2}, \ \forall t \geq t'.$$

8.3 Proof of Theorem 2

First, we show that the inflation factor in the UCB estimate is less than some constant $\epsilon > 0$ with probability greater than $1 - 1/t^2$.

Lemma 3 *Given $\eta \in (0,1)$ $\kappa > 0$ and $\epsilon > 0$, for all integers $t \geq \max\{t_0' := 4c_0^2/e^2, 4c^2/e^2\}$ where $c_0 := 3\kappa^2/(2\epsilon^2 p^* \eta)$ and $c := 1/(p^*(1-\eta))^2$, we have*

$$\Pr\left(\kappa\sqrt{\frac{3 \ln t}{2T_i^{t+1}}} \geq \epsilon\right) \leq \frac{1}{t^2}.$$

Proof. Note that

$$\kappa\sqrt{\frac{3\ln t}{2T_i^{t+1}}} \geq \epsilon \iff \kappa^2 \frac{3\ln t}{2T_i^{t+1}} \geq \epsilon^2 \iff T_i^{t+1} \leq \frac{3\kappa^2 \ln t}{2\epsilon^2} \iff T_i^{t+1} \leq \frac{3\kappa^2 \ln t}{2\epsilon^2 p^* t} p^* t.$$

Hence, when $\frac{3\kappa^2 \ln t}{2\epsilon^2 p^* t} \leq \eta$, we have

$$\left\{\kappa\sqrt{\frac{3\ln t}{2T_i^{t+1}}} \geq \epsilon\right\} \Rightarrow \{T_i^{t+1} \leq \eta p^* t\}$$

which implies that for $t \geq t' = 4c^2/\epsilon^2$,

$$\Pr\left(\kappa\sqrt{\frac{3\ln t}{2T_i^{t+1}}} \geq \epsilon\right) \leq \Pr(T_i^{t+1} \leq \eta p^* t) \leq \frac{1}{t^2} \tag{6}$$

by Lemma 2. Next, following the procedure we use in the proof of Lemma 2, we show that the inequality $\frac{3\kappa^2 \ln t}{2\epsilon^2 p^* t} \leq \eta$ is satisfied when $t \geq t_0'$, which will imply that the result in (6) is satisfied for $t \geq \max\{t', t_0'\}$.

We want to show that, given some $\eta \in (0,1)$, for t sufficiently large, the inequality $\frac{3\kappa^2 \ln t}{2\epsilon^2 p^* t} \leq \eta$ holds. Let $h_0(t) := 2\epsilon^2 p^* \eta t$ and $g_0(t) := 3\kappa^2 \ln t$. Let $t_0^+ \in \mathbb{R}$ be the greatest root of the equation $h_0(t) = g_0(t)$, i.e., for all $t > t_0^+$, $h_0(t) > g_0(t)$. To see this, we observe that $h_0(1) > g_0(1)$, and $h_0'(t) = 2\epsilon^2 p^* \eta$ and $g_0'(t) = 3\kappa^2/t$. Therefore $\lim_{t\to\infty} g_0'(t) = 0$ whereas $h_0'(t) = 2\epsilon^2 p^* \eta \in \mathbb{R}^+$ for all $t \in \mathbb{N}_+$. As a result, $\exists t'' \in \mathbb{R}$ such that $h_0'(t) > g_0'(t)$, $\forall t \geq t''$. This justifies that t_0^+ is finite.

Next, we find an upper bound for t_0^+. Note that $h_0(t) = g_0(t)$ implies that $\ln t = 2t\epsilon^2 p^* \eta/(3\kappa^2)$. Let $c_0 = 3\kappa^2/(2\epsilon^2 p^* \eta)$. Thus, taking exponentials of the both sides, we want to solve the equation $e^{t/c_0} = t$ whose roots are given by the equation

$$t = \frac{c_0}{\ln e}\text{glog}\left(\frac{c_0}{\ln e}\right) = c_0 \text{glog}(c_0).$$

Then, using the bound given in Kalman (2001), we obtain

$$t_0^+ \leq c_0^2 \left(\frac{k}{e}\right)^{\frac{k}{k-1}} \quad \text{for } k > 1.$$

By setting $k = 2$, we obtain the following upper bound on t_0^+:

$$t_0^+ \leq \frac{4c_0^2}{e^2}. \tag{7}$$

This implies that $\forall t \geq 4c_0^2/e^2 \geq t_0^+$, we have $h_0(t) \geq g_0(t)$, which further implies that $\frac{3\kappa^2 \ln t}{2\epsilon^2 p^* t} \leq \eta$. Hence, the inequalities in (6) hold for all integers $t \geq \max\{4c_0^2/e^2, 4c^2/e^2\}$. □

The next lemma shows that the sample mean estimate of the expected state of any arm cannot be too far away from its true value when t is large.

Lemma 4 *For any $\delta > 0$ and $\eta \in (0,1)$, we have*

$$\Pr(|\hat{\mu}_i^{t+1} - \mu_i| \geq \delta) \leq \frac{2}{t^2(1-e^{-2\delta^2})} + 2e^{-2\delta^2 \eta p^* t}$$

for all integers $t \geq t' := 4c^2/e^2$, where $c = 1/(p^*(1-\eta))^2$.

Proof. Given that $T_i^{t+1} = j$, we have $j\hat{\mu}_i^{t+1} = \sum_{k=1}^t X_i^{(k)} 1_{\{i \in \tau_k\}}$. Since $X_i^{(k)} 1_{\{i\in\tau_k\}} \in [0,1]$, by Hoeffding's inequality, we obtain

$$\begin{aligned}
\Pr(|\mu_i - \hat{\mu}_i^{t+1}| \geq \delta | T_i^{t+1} = j) &= \Pr(|T_i^{t+1}\mu_i - T_i^{t+1}\hat{\mu}_i^{t+1}| \geq T_i^{t+1}\delta | T_i^{t+1} = j) \\
&= \Pr(|\mathbb{E}[T_i^{t+1}\hat{\mu}_i^{t+1}] - T_i^{t+1}\hat{\mu}_i^{t+1}| \geq j\delta | T_i^{t+1} = j) \\
&= \Pr(|\mathbb{E}[j\hat{\mu}_i^{t+1}] - j\hat{\mu}_i^{t+1}| \geq j\delta | T_i^{t+1} = j) \\
&\leq 2e^{-2(\delta j)^2/j} = 2e^{-2\delta^2 j}.
\end{aligned}$$

Then, using the law of total probability, we obtain

$$\begin{aligned}
\Pr(|\mu_i - \hat{\mu}_i^{t+1}| \geq \delta) &= \sum_{j=0}^t \Pr(|\mu_i - \hat{\mu}_i^{t+1}| \geq \delta | T_i^{t+1} = j) \Pr(T_i^{t+1} = j) \\
&\leq \sum_{j=0}^t 2e^{-2\delta^2 j} \Pr(T_i^{t+1} = j).
\end{aligned}$$

Let $t^* = \eta p^* t$. For $0 \leq j < t^*$, we have $\{T_i^{t+1} = j\} \subseteq \{T_i^{t+1} \leq t^*\}$; hence, $\Pr(T_i^{t+1} = j) \leq \Pr(T_i^{t+1} \leq t^*) \leq 1/t^2$ when $t \geq t'$ by Lemma 2. We also have $\sum_{j=\lceil t^* \rceil}^t \Pr(T_i^{t+1} = j) \leq 1$. Using these, we proceed as follows: For $t \geq t'$, we have

$$\begin{aligned}
\Pr(|\mu_i - \hat{\mu}_i^{t+1}| \geq \delta) &\leq \sum_{j=0}^{\lfloor t^* \rfloor} 2e^{-2\delta^2 j} \Pr(T_i^{t+1} = j) + \sum_{j=\lceil t^* \rceil}^t 2e^{-2\delta^2 j} \Pr(T_i^{t+1} = j) \\
&\leq \frac{1}{t^2} \sum_{j=0}^{\lfloor t^* \rfloor} 2e^{-2\delta^2 j} + 2e^{-2\delta^2 \lceil t^* \rceil} \sum_{j=\lceil t^* \rceil}^t \Pr(T_i^{t+1} = j) \\
&\leq \frac{2}{t^2(1-e^{-2\delta^2})} + 2e^{-2\delta^2 \lceil t^* \rceil} \\
&\leq \frac{2}{t^2(1-e^{-2\delta^2})} + 2e^{-2\delta^2 \eta p^* t}.
\end{aligned}$$

□

Note that $\kappa\sqrt{\frac{3\ln t}{2T_i^{t+1}}} < \delta/2$ and $|\hat{\mu}_i^{t+1} - \mu_i| < \delta/2$ for all $i \in \{1,\ldots,m\}$ implies that $\Delta_{t+1}^\kappa < \delta$ by the triangle inequality. Hence, by the contrapositive of this proposition, we have $\{\Delta_{t+1}^\kappa \geq \delta\} \subseteq \left\{\bigcup_i \{|\hat{\mu}_i^{t+1} - \mu_i| \geq \delta/2\}\right\} \cup \left\{\bigcup_i \{\kappa\sqrt{\frac{3\ln t}{2T_i^{t+1}}} \geq \delta/2\}\right\}$. Using the union bound and the results of Lemmas 3 and 4, we obtain

$$\begin{aligned}
\Pr(\Delta_{t+1}^\kappa \geq \delta) &\leq \sum_{i=1}^m \Pr(|\hat{\mu}_i^{t+1} - \mu_i| \geq \delta/2) + \sum_{i=1}^m \Pr\left(\kappa\sqrt{\frac{3\ln t}{2T_i^{t+1}}} \geq \delta/2\right) \\
&\leq \frac{2m}{t^2(1-e^{-\delta^2/2})} + 2me^{-\delta^2 \eta p^* t/2} + \frac{m}{t^2}
\end{aligned}$$

for $t \geq t_1$. When $\kappa = 0$, the result directly follows from Lemma 4.

8.4 Proof of Theorem 3

Fix any $\eta \in (0,1)$. Note that $\delta > 0$ since $\nabla_{\min} > 0$. First, we prove the theorem for $\kappa > 0$. For $T \leq \lceil t_1 \rceil$, the regret bound is $T\nabla_{\max}$. By Lemma 1, we have

$$\begin{aligned}\{\Delta_t^\kappa < \delta\} &\Rightarrow \Pr(r_{\boldsymbol{\mu}}(S_t) \geq \alpha r_{\boldsymbol{\mu}}^*) \geq \beta \\ &\Rightarrow \mathbb{E}[r_{\boldsymbol{\mu}}(S_t)] \geq \alpha\beta r_{\boldsymbol{\mu}}^* \Rightarrow \mathcal{N}_t \\ &\Rightarrow \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)] \leq 0.\end{aligned}$$

Hence, we have the following for $T > \lceil t_1 \rceil$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)\right] = \mathbb{E}\left[\sum_{t=1}^{\lceil t_1 \rceil} \mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)\right] + \mathbb{E}\left[\sum_{t=\lceil t_1 \rceil+1}^{T} \mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)\right] \\
&\leq \nabla_{\max}\lceil t_1 \rceil + \sum_{t=\lceil t_1 \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)|\Delta_t^\kappa < \delta] \Pr(\Delta_t^\kappa < \delta) \\
&+ \sum_{t=\lceil t_1 \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)|\Delta_t^\kappa \geq \delta] \Pr(\Delta_t^\kappa \geq \delta) \\
&= \nabla_{\max}\lceil t_1 \rceil + \sum_{t=\lceil t_1 \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)|\mathcal{N}_t] \Pr(\Delta_t^\kappa < \delta) \\
&+ \sum_{t=\lceil t_1 \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)|\Delta_t^\kappa \geq \delta] \Pr(\Delta_t^\kappa \geq \delta) \\
&\leq \nabla_{\max}\lceil t_1 \rceil + \sum_{t=\lceil t_1 \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)|\Delta_t^\kappa \geq \delta] \\
&\times \left(\frac{2m}{(t-1)^2(1-e^{-\delta^2/2})} + 2me^{-\delta^2\eta p^*(t-1)/2} + \frac{m}{(t-1)^2}\right) \\
&\leq \nabla_{\max}\lceil t_1 \rceil + \nabla_{\max}\sum_{t=1}^{\infty}\left(\frac{2m}{t^2(1-e^{-\delta^2/2})} + 2me^{-\delta^2\eta p^* t/2} + \frac{m}{t^2}\right) \\
&\leq \nabla_{\max}\left(\lceil t_1 \rceil + \frac{m\pi^2}{3(1-e^{-\delta^2/2})} + \frac{2me^{-\delta^2\eta p^*/2}}{1-e^{-\delta^2\eta p^*/2}} + \frac{m\pi^2}{6}\right) \\
&\leq \nabla_{\max}\left(\lceil t_1 \rceil + \frac{m\pi^2}{3(1-e^{-\delta^2/2})} + \frac{2m}{1-e^{-\delta^2\eta p^*/2}} + \frac{m\pi^2}{6}\right).
\end{aligned}$$

We know that $\frac{1}{1-e^{-x}} \leq 1 + \frac{1}{x}$ for $x > 0$. Using this, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\text{CUCB-}\kappa}(t)\right] &\leq \nabla_{\max}\left(\lceil t_1 \rceil + \frac{m\pi^2}{3(1-e^{-\delta^2/2})} + \frac{2m}{1-e^{-\delta^2\eta p^*/2}} + \frac{m\pi^2}{6}\right) \\
&\leq \nabla_{\max}\left(\lceil t_1 \rceil + \frac{m\pi^2}{3}\left(1 + \frac{2}{\delta^2} + \frac{1}{2}\right) + 2m\left(1 + \frac{2}{\delta^2\eta p^*}\right)\right).
\end{aligned} \tag{8}$$

Since (8) holds for any $\eta \in (0,1)$, the final result is obtained by minimizing (8) over $\eta \in (0,1)$.

For $\kappa = 0$, we follow the same procedure. For $T \leq \lceil t' \rceil$, the regret bound is $T\nabla_{\max}$. By Lemma 1, we have

$$\begin{aligned}\{\Delta_t < \delta\} &\Rightarrow \Pr(r_{\boldsymbol{\mu}}(S_t) \geq \alpha r^*_{\boldsymbol{\mu}}) \geq \beta \\ &\Rightarrow \mathbb{E}[r_{\boldsymbol{\mu}}(S_t)] \geq \alpha\beta r^*_{\boldsymbol{\mu}} \Rightarrow \mathcal{N}_t \\ &\Rightarrow \mathbb{E}[\mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t)] \leq 0.\end{aligned}$$

Hence, we have the following for $T > \lceil t' \rceil$:

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t)\right] = \mathbb{E}\left[\sum_{t=1}^{\lceil t' \rceil} \mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t)\right] + \mathbb{E}\left[\sum_{t=\lceil t' \rceil+1}^{T} \mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t)\right] \\ &\leq \nabla_{\max}\lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}\left[\mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t) | \Delta_t < \delta\right] \Pr(\Delta_t < \delta) \\ &+ \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}\left[\mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t) | \Delta_t \geq \delta\right] \Pr(\Delta_t \geq \delta) \\ &\leq \nabla_{\max}\lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}\left[\mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t) | \Delta_t \geq \delta\right] \\ &\times \left(\frac{2m}{(t-1)^2(1-e^{-2\delta^2})} + 2me^{-2\delta^2\eta p^*(t-1)}\right) \\ &\leq \nabla_{\max}\lceil t' \rceil + \nabla_{\max}\sum_{t=1}^{\infty}\left(\frac{2m}{t^2(1-e^{-2\delta^2})} + 2me^{-2\delta^2\eta p^* t}\right) \\ &\leq \nabla_{\max}\left(\lceil t' \rceil + \frac{m\pi^2}{3(1-e^{-2\delta^2})} + \frac{2m}{1-e^{-2\delta^2\eta p^*}}\right).\end{aligned}$$

We know that $\frac{1}{1-e^{-x}} \leq 1 + \frac{1}{x}$ for $x > 0$. Using this, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t)\right] &\leq \nabla_{\max}\left(\lceil t' \rceil + \frac{m\pi^2}{3(1-e^{-2\delta^2})} + \frac{2m}{1-e^{-2\delta^2\eta p^*}}\right) \\ &\leq \nabla_{\max}\left(\lceil t' \rceil + \frac{m\pi^2}{3}\left(1 + \frac{1}{2\delta^2}\right) + 2m\left(1 + \frac{1}{2\delta^2\eta p^*}\right)\right). \quad (9)\end{aligned}$$

Finally, we minimize (9) over $\eta \in (0,1)$ via the procedure we follow for minimizing (8).

8.5 Proof of Theorem 4

For $T \leq \lceil t' \rceil$, the regret bound is $T\nabla_{\max}$. To establish a regret bound for $T > \lceil t' \rceil$, first, we show that the instantaneous regret $\mathrm{IR}^{\text{CUCB-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t)$ can be bounded by using Δ^0_t and the bounded smoothness function f.

Lemma 5 *For any $t \in \mathbb{N}_+$ and expectation vector $\boldsymbol{\mu}$, the instantaneous regret of CUCB-κ is bounded as follows:*

$$\mathbb{E}[IR^{CUCB\text{-}\kappa}_{\boldsymbol{\mu},\alpha,\beta}(t)] \leq 2\mathbb{E}[f(\Delta^{\kappa}_t)].$$

Proof. Taking expectation of both sides of the equation (4) in Lemma 1, we obtain the desired result. □

In order to make use of Lemma 5 and find an upper bound for the instantaneous regret of CUCB-0, we bound the value $\mathbb{E}[\Delta^0_{t+1}]$ for $t > \lceil t' \rceil$.

Lemma 6 *For any $\eta \in (0,1)$, we have*

$$\mathbb{E}[\Delta^0_{t+1}] \leq t^{-1/2}\left(2m\sqrt{\frac{\pi}{2\eta p^*}} + 3m\right)$$

for all integers $t \geq t' := 4c^2/e^2$, where $c := 1/(p^*(1-\eta))^2$.

Proof. We have

$$\mathbb{E}[\Delta^0_{t+1}] = \int_0^1 \Pr(\Delta^0_{t+1} \geq \delta)\mathrm{d}\delta.$$

Hence, by Theorem 2, since $\Delta^0_{t+1} \in [0,1]$, we have for any $\theta > 0$

$$\begin{aligned}\mathbb{E}[\Delta^0_{t+1}] &\leq \int_0^{t^{-\theta}} \Pr(\Delta^0_{t+1} \geq \delta)\mathrm{d}\delta + \int_{t^{-\theta}}^1 \left(\frac{2m}{t^2(1-e^{-2\delta^2})} + 2me^{-2\delta^2\eta p^* t}\right)\mathrm{d}\delta \\ &\leq t^{-\theta} + \int_{t^{-\theta}}^1 \left(\frac{2m}{t^2(1-e^{-2\delta^2})} + 2me^{-2\delta^2\eta p^* t}\right)\mathrm{d}\delta\end{aligned}$$

for $t > \lceil t' \rceil$.

Note that $e^{-2\delta^2\eta p^* t}$ is just a scaled version of the pdf of the normal distribution. Let $B = 1/(2\sqrt{\eta p^* t})$, then we have

$$\int e^{-2\delta^2\eta p^* t}\mathrm{d}\delta = \int e^{-\delta^2/(2B^2)}\mathrm{d}\delta = B\sqrt{2\pi}\int \frac{1}{B\sqrt{2\pi}} e^{\frac{-\delta^2}{2B^2}}\mathrm{d}\delta.$$

Hence, setting $B = \sqrt{1/(2At)}$ ($A = 2\eta p^*$), via the fact that $1/(1-e^{-x}) \leq 1 + 1/x$ for all $x > 0$, we obtain

$$\begin{aligned}\mathbb{E}[\Delta^0_{t+1}] &\leq t^{-\theta} + 2m\sqrt{\frac{\pi}{At}} + \int_{t^{-\theta}}^1 \left(\frac{2m}{t^2(1-e^{-2\delta^2})}\right)\mathrm{d}\delta \\ &\leq t^{-\theta} + 2m\sqrt{\frac{\pi}{At}} + \frac{m}{t^2}\int_{t^{-\theta}}^1 \left(2 + \frac{1}{\delta^2}\right)\mathrm{d}\delta \\ &= t^{-\theta} + 2m\sqrt{\frac{\pi}{At}} + mt^{-2}\left(2\delta - \frac{1}{\delta}\right)\Bigg|_{\delta=t^{-\theta}}^{\delta=1} \\ &= t^{-\theta} + 2m\sqrt{\frac{\pi}{At}} + mt^{-2}\left(2 - 1 - 2t^{-\theta} + t^{\theta}\right) \\ &\leq t^{-\theta} + 2m\sqrt{\frac{\pi}{At}} + mt^{-2}(t^{\theta} + 1)\end{aligned}$$

$$= t^{-\theta} + 2m\sqrt{\frac{\pi}{At}} + mt^{\theta-2} + mt^{-2}.$$

The order of t in this bound is minimized when $\theta = 1$; thus, noticing that $2m+1 \leq 3m$, we have

$$\begin{aligned}
\mathbb{E}[\Delta^0_{t+1}] &\leq 2m\sqrt{\frac{\pi}{At}} + (m+1)t^{-1} + mt^{-2} \\
&\leq 2mt^{-1/2}\sqrt{\frac{\pi}{A}} + (m+1)t^{-1/2} + mt^{-1/2} \\
&\leq t^{-1/2}\Big(2m\sqrt{\frac{\pi}{A}} + 2m + 1\Big) \\
&\leq t^{-1/2}\Big(2m\sqrt{\frac{\pi}{A}} + 3m\Big).
\end{aligned}$$

□

Next, using this lemma (Lemma 6), we show that the gap-independent regret for CUCB-κ (when $\kappa = 0$) is at most $O(T^{1-w/2})$ for a large set of problems where the bounded-smoothness function $f(x)$ is γx^{ω}, for $\omega \in (0, 1]$.

Note that f is concave. Hence, via Lemma 5, Lemma 6 and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{IR}^{\text{CUCB-0}}_{\boldsymbol{\mu},\alpha,\beta}(t+1)] &\leq 2\mathbb{E}[f(\Delta^0_{t+1})] \\
&\leq 2f(\mathbb{E}[\Delta^0_{t+1}]) \\
&\leq 2f\left(2m\sqrt{\frac{\pi}{At}} + 3m\sqrt{\frac{1}{t}}\right) \\
&\leq f\left(4m\sqrt{\frac{\pi}{At}}\right) + f\left(6m\sqrt{\frac{1}{t}}\right) \\
&= \gamma\left((4m)^{\omega}\left(\frac{\pi}{At}\right)^{\omega/2} + (6m)^{\omega}t^{-\omega/2}\right).
\end{aligned}$$

Now, we can bound the regret by bounding the sum of expected instantaneous regrets throughout the time horizon:

$$\begin{aligned}
Reg^{\text{CUCB-0}}_{\boldsymbol{\mu},\alpha,\beta}(T) &= \sum_{t=1}^{T}\mathbb{E}[\mathrm{IR}^{\text{CUCB-0}}_{\boldsymbol{\mu},\alpha,\beta}(t)] \\
&\leq \sum_{t=1}^{\lceil t'\rceil}\mathbb{E}[\mathrm{IR}^{\text{CUCB-0}}_{\boldsymbol{\mu},\alpha,\beta}(t)] + \sum_{t=\lceil t'\rceil+1}^{T}\mathbb{E}[\mathrm{IR}^{\text{CUCB-0}}_{\boldsymbol{\mu},\alpha,\beta}(t)] \\
&\leq \lceil t'\rceil\Delta_{\max} + \gamma(2m)^{\omega}\left[2^{\omega}\left(\frac{\pi}{A}\right)^{\omega/2} + 3^{\omega}\right]\sum_{t=1}^{T}t^{-\omega/2}.
\end{aligned}$$

Using the fact that $\sum_{t=1}^{T} t^{-x} \leq T^{1-x}/(1-x)\ \forall x \in (0,1)$, we have the following for $\omega \in (0, 1]$:

$$Reg^{\text{CUCB-0}}_{\boldsymbol{\mu},\alpha,\beta}(T) \leq \lceil t'\rceil\nabla_{\max} + \gamma(2m)^{\omega}\left[2^{\omega}\left(\frac{\pi}{A}\right)^{\omega/2} + 3^{\omega}\right]\frac{T^{1-w/2}}{1-w/2}.$$

Finally, since the above regret bound holds for any $\eta \in (0,1)$, we take the infimum over η.

8.6 Proof of Theorem 5

First, we let $F_{n,p}^{\text{Binom}}(x)$ be the cdf of Binomial(n,p) and $F_{\alpha,\beta}^{\text{Beta}}(x)$ be the cdf of Beta(α,β). Then, we define a series of lemmas and facts that will be used in the proof.

Lemma 7 *(Lemma 9 in Agrawal and Goyal (2012))* $F_{n,p}^{Binom}(np - n\delta) \leq e^{-2n\delta^2}$, $1 - F_{n,p}^{Binom}(np + n\delta) \leq e^{-2n\delta^2}$ *and* $1 - F_{n+1,p}^{Binom}(np + n\delta) \leq e^{4\delta - 2n\delta^2}$ *for all* $n \in \mathbb{N}_+$, $p \in [0,1]$, $\delta \geq 0$.

Fact 1. $F_{n+1,p}^{Binom}(i) \geq F_{n,p}^{Binom}(i-1)$ *for all* $n \in \mathbb{N}_+$, $p \in [0,1]$, *and integers* i *such that* $1 \leq i \leq n$.

Proof.

$$F_{n+1,p}^{\text{Binom}}(i) \geq (1-p)F_{n,p}^{\text{Binom}}(i) + pF_{n,p}^{\text{Binom}}(i-1) \geq F_{n,p}^{\text{Binom}}(i-1)$$

□

Fact 2. *(Fact 1 in Agrawal and Goyal (2012))*

$$F_{\alpha,\beta}^{Beta}(y) = 1 - F_{\alpha+\beta-1,y}^{Binom}(\alpha - 1)$$

for all $\alpha, \beta \in \mathbb{N}_+$ *and* $y \in [0,1]$.

Next, as in the CUCB-κ case, we prove that ν_i^{t+1} is lower than some constant with high probability for t sufficiently large. Recalling that $\Delta_{t+1}^0 = |\hat{\mu}_i^{t+1} - \mu_i|$, we have the following by the law of total probability:

$$\begin{aligned}
&\Pr(|\nu_i^{t+1} - \mu_i| \geq \delta) \\
&= \Pr(|\nu_i^{t+1} - \mu_i| \geq \delta, \Delta_{t+1}^0 \geq \delta/2) + \Pr(|\nu_i^{t+1} - \mu_i| \geq \delta, \Delta_{t+1}^0 < \delta/2) \\
&\leq \Pr(\Delta_{t+1}^0 \geq \delta/2) + \Pr(|\nu_i^{t+1} - \mu_i| \geq \delta, \Delta_{t+1}^0 < \delta/2). \qquad (10)
\end{aligned}$$

By Lemma 4, we already have an upper bound for $\Pr(\Delta_{t+1}^0 \geq \delta/2)$ for $t \geq t'$. Note that $|\nu_i^{t+1} - \mu_i| \geq \delta$ and $\Delta_{t+1}^0 < \delta/2$ implies that $|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2$ by the triangle inequality. Hence, we have

$$\Pr(\Delta_{t+1}^0 < \delta/2, |\nu_i^{t+1} - \mu_i| \geq \delta) \leq \Pr(|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2). \qquad (11)$$

Now, we will find an upper bound for the RHS of this inequality. First, notice that we have the following

$$\begin{aligned}
&\Pr(|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2) \\
&= \sum_{j=0}^{t} \Pr(|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2 | T_i^{t+1} = j) \Pr(T_i^{t+1} = j) \\
&\leq \sum_{j=0}^{t} \Pr(\nu_i^{t+1} - \hat{\mu}_i^{t+1} \geq \delta/2 | T_i^{t+1} = j) \Pr(T_i^{t+1} = j)
\end{aligned}$$

$$+\sum_{j=0}^{t}\Pr(\nu_i^{t+1}-\hat{\mu}_i^{t+1}\le -\delta/2|T_i^{t+1}=j)\Pr(T_i^{t+1}=j). \tag{12}$$

In the expression below, the second equality comes from the Fact 2 and the last inequality is obtained via the Lemma 7:

$$\begin{aligned}\Pr(\nu_i^{t+1}-\hat{\mu}_i^{t+1}\ge \delta/2|T_i^{t+1}=j) &= \mathbb{E}[1-F^{\text{Beta}}_{1+s_i^{t+1},j-s_i^{t+1}+1}(\hat{\mu}_i^{t+1}+\delta/2)|T_i^{t+1}=j]\\ &= \mathbb{E}[F^{\text{Binom}}_{j+1,\hat{\mu}_i^{t+1}+\delta/2}(s_i^{t+1})|T_i^{t+1}=j]\\ &\le \mathbb{E}[F^{\text{Binom}}_{j,\hat{\mu}_i^{t+1}+\delta/2}(s_i^{t+1})|T_i^{t+1}=j]\\ &= \mathbb{E}[F^{\text{Binom}}_{j,\hat{\mu}_i^{t+1}+\delta/2}(j\hat{\mu}_i^{t+1})|T_i^{t+1}=j]\\ &\le e^{-\delta^2 j/2}. \end{aligned}\tag{13}$$

Similarly, by Fact 2 and Lemma 7, we also have

$$\begin{aligned}\Pr(\nu_i^{t+1}-\hat{\mu}_i^{t+1}\le -\delta/2|T_i^{t+1}=j) &= \mathbb{E}[F^{\text{Beta}}_{1+s_i^{t+1},j-s_i^{t+1}+1}(\hat{\mu}_i^{t+1}-\delta/2)|T_i^{t+1}=j]\\ &= \mathbb{E}[1-F^{\text{Binom}}_{j+1,\hat{\mu}_i^{t+1}-\delta/2}(s_i^{t+1})|T_i^{t+1}=j]\\ &= \mathbb{E}[1-F^{\text{Binom}}_{j+1,\hat{\mu}_i^{t+1}-\delta/2}(j\hat{\mu}_i^{t+1})|T_i^{t+1}=j]\\ &\le e^{2\delta-\delta^2 j/2}. \end{aligned}\tag{14}$$

Using (13) and (14) in (12), we obtain

$$\Pr(|\nu_i^{t+1}-\hat{\mu}_i^{t+1}|\ge\delta/2)\le(e^{2\delta}+1)\sum_{j=0}^{t}e^{-\delta^2 j/2}\Pr(T_i^{t+1}=j).$$

For $0\le j<t^*=\eta p^* t$, we have $\Pr(T_i^{t+1}=j)\le\Pr(T_i^{t+1}\le t^*)\le 1/t^2$ when $t\ge t'$ by Lemma 2. We also have $\sum_{j=\lceil t^*\rceil}^{t}\Pr(T_i^{t+1}=j)\le 1$. Using these, we proceed as follows: For $t\ge t'$, we have

$$\begin{aligned}&\frac{\Pr(|\nu_i^{t+1}-\mu_i|\ge\delta,\Delta_{t+1}^0<\delta/2)}{e^{2\delta}+1}\\ &\le\frac{\Pr(|\nu_i^{t+1}-\hat{\mu}_i^{t+1}|\ge\delta/2)}{e^{2\delta}+1}\\ &\le\sum_{j=0}^{\lfloor t^*\rfloor}e^{-\delta^2 j/2}\Pr(T_i^{t+1}=j)+\sum_{j=\lceil t^*\rceil}^{t}e^{-\delta^2 j/2}\Pr(T_i^{t+1}=j)\\ &\le\frac{1}{t^2}\sum_{j=0}^{\lfloor t^*\rfloor}e^{-\delta^2 j/2}+e^{-\delta^2\lceil t^*\rceil/2}\sum_{j=\lceil t^*\rceil}^{t}\Pr(T_i^{t+1}=j)\\ &\le\frac{1}{t^2(1-e^{-\delta^2/2})}+e^{-\delta^2\lceil t^*\rceil/2}\\ &\le\frac{1}{t^2(1-e^{-\delta^2/2})}+e^{-\delta^2\eta p^* t/2}. \end{aligned}\tag{15}$$

Using the result of Lemma 4 and (15) in (10), and recalling the inequality in (11), we obtain

$$\Pr(|\nu_i^{t+1} - \mu_i| \geq \delta) \leq (3 + e^{2\delta})\Big[\frac{1}{t^2(1 - e^{-\delta^2/2})} + e^{-\delta^2 \eta p^* t/2}\Big]$$

for $t \geq t'$, and for any $i \in \{1, \ldots, m\}$ and $\eta \in (0, 1)$.

Finally by using the union bound, we obtain

$$\Pr\left(\bigcup_i \{|\nu_i^{t+1} - \mu_i| \geq \delta\}\right) \leq (3 + e^{2\delta})m\Big[\frac{1}{t^2(1 - e^{-\delta^2/2})} + e^{-\delta^2 \eta p^* t/2}\Big]$$

for $t \geq t'$.

8.7 Proof of Theorem 6

We follow the procedure we use in Theorem 3. Fix any $\eta \in (0, 1)$. Note that $\delta > 0$ since $\nabla_{\min} > 0$. For $T \leq \lceil t' \rceil$, the regret bound is $T\nabla_{\max}$. Recall that $\Delta_t^{\boldsymbol{\nu}} = \max_i |\nu_i^t - \mu_i|$. By Lemma 1, we have

$$\begin{aligned}
\{\Delta_t^{\boldsymbol{\nu}} < \delta\} &\Rightarrow \Pr(r_{\boldsymbol{\mu}}(S_t) \geq \alpha r_{\boldsymbol{\mu}}^*) \geq \beta \\
&\Rightarrow \mathbb{E}[r_{\boldsymbol{\mu}}(S_t)] \geq \alpha\beta r_{\boldsymbol{\mu}}^* \Rightarrow \mathcal{N}_t \\
&\Rightarrow \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)] \leq 0.
\end{aligned}$$

Hence, we have the following for $T > \lceil t' \rceil$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)\right] = \mathbb{E}\left[\sum_{t=1}^{\lceil t' \rceil} \mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)\right] + \mathbb{E}\left[\sum_{t=\lceil t' \rceil+1}^{T} \mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)\right] \\
&\leq \nabla_{\max}\lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)|\Delta_t^{\boldsymbol{\nu}} < \delta] \Pr(\Delta_t^{\boldsymbol{\nu}} < \delta) \\
&+ \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)|\Delta_t^{\boldsymbol{\nu}} \geq \delta] \Pr(\Delta_t^{\boldsymbol{\nu}} \geq \delta) \\
&= \nabla_{\max}\lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)|\mathcal{N}_t] \Pr(\Delta_t^{\boldsymbol{\nu}} < \delta) \\
&+ \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)|\Delta_t^{\boldsymbol{\nu}} \geq \delta] \Pr(\Delta_t^{\boldsymbol{\nu}} \geq \delta) \\
&\leq \nabla_{\max}\lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}[\mathrm{IR}_{\boldsymbol{\mu},\alpha,\beta}^{\mathrm{CTS}}(t)|\Delta_t^{\boldsymbol{\nu}} \geq \delta] \\
&\times \left(\frac{(3 + e^{2\delta})m}{(t-1)^2(1 - e^{-\delta^2/2})} + (3 + e^{2\delta})me^{-\delta^2 \eta p^*(t-1)/2}\right) \\
&\leq \nabla_{\max}\lceil t' \rceil + \nabla_{\max}\sum_{t=1}^{\infty}\left(\frac{(3 + e^{2\delta})m}{t^2(1 - e^{-\delta^2/2})} + (3 + e^{2\delta})me^{-\delta^2 \eta p^* t/2}\right)
\end{aligned}$$

$$\leq \nabla_{\max}\left(\lceil t'\rceil + \frac{(3+e^{2\delta})m\pi^2}{6(1-e^{-\delta^2/2})} + \frac{(3+e^{2\delta})me^{-\delta^2\eta p^*/2}}{1-e^{-\delta^2\eta p^*/2}}\right)$$
$$\leq \nabla_{\max}\left(\lceil t'\rceil + \frac{(3+e^{2\delta})m\pi^2}{6(1-e^{-\delta^2/2})} + \frac{(3+e^{2\delta})m}{1-e^{-\delta^2\eta p^*/2}}\right).$$

We know that $\frac{1}{1-e^{-x}} \leq 1+\frac{1}{x}$ for $x>0$. Using this, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathrm{IR}^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(t)\right] \leq \nabla_{\max}\left(\lceil t'\rceil + \frac{(3+e^{2\delta})m\pi^2}{6}\left(1+\frac{2}{\delta^2}\right) + (3+e^{2\delta})m\left(1+\frac{2}{\delta^2\eta p^*}\right)\right).$$

Finally, since the regret bound above holds for any $\eta \in (0,1)$, we take the infimum over η.

8.8 Proof of Theorem 7

For $T \leq \lceil t'\rceil$, the regret bound is $T\nabla_{\max}$. To find the regret bound for $T > \lceil t'\rceil$, first, we show that the instantaneous regret $\mathrm{IR}^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(t)$ can be bounded by using $\Delta^{\boldsymbol{\nu}}_t$ and the bounded smoothness function f.

Lemma 8 *For any integer $t \geq 1$, expectation vector $\boldsymbol{\mu}$, the instantaneous regret of CTS is bounded as follows:*

$$\mathbb{E}[IR^{CTS}_{\boldsymbol{\mu},\alpha,\beta}(t)] \leq 2\mathbb{E}[f(\Delta^{\boldsymbol{\nu}}_t)].$$

Proof. Taking expectation of both sides of the equation (4) in Lemma 1, we obtain the desired result. □

The next step is to bound $\mathbb{E}[\Delta^{\boldsymbol{\nu}}_{t+1}]$ for $t > \lceil t'\rceil$.

Lemma 9 *For any $\eta \in (0,1)$, we have*

$$\mathbb{E}[\Delta^{\boldsymbol{\nu}}_{t+1}] \leq (3+e^2)mt^{-1/2}\Big(\sqrt{\frac{2\pi}{\eta p^*}}+3\Big)$$

for all integers $t \geq t' := 4c^2/e^2$, where $c := 1/(p^*(1-\eta))^2$.

Proof. We have

$$\mathbb{E}[\Delta^{\boldsymbol{\nu}}_{t+1}] = \int_0^1 \Pr(\Delta^{\boldsymbol{\nu}}_{t+1} \geq \delta)\mathrm{d}\delta.$$

Hence, by Theorem 5, since $\Delta^{\boldsymbol{\nu}}_{t+1} \in [0,1]$, we have for any $\theta > 0$

$$\mathbb{E}[\Delta^{\boldsymbol{\nu}}_{t+1}] \leq \int_0^{t^{-\theta}} \Pr(\Delta^{\boldsymbol{\nu}}_{t+1} \geq \delta)\mathrm{d}\delta + \int_{t^{-\theta}}^1 \Big(\frac{(3+e^{2\delta})m}{t^2(1-e^{-\delta^2/2})} + (3+e^{2\delta})me^{-\delta^2\eta p^* t/2}\Big)\mathrm{d}\delta$$
$$\leq t^{-\theta} + \int_{t^{-\theta}}^1 \Big(\frac{(3+e^2)m}{t^2(1-e^{-\delta^2/2})} + (3+e^2)me^{-\delta^2\eta p^* t/2}\Big)\mathrm{d}\delta$$

for $t > \lceil t'\rceil$.

Note that $e^{-\delta^2 \eta p^* t/2}$ is just a scaled version of the probability distribution function of normal distribution. Let $B = \sqrt{1/(\eta p^* t)}$, then we have

$$\int e^{-\delta^2 \eta p^* t/2} \mathrm{d}\delta = \int e^{-\delta^2/(2B^2)} \mathrm{d}\delta = B\sqrt{2\pi} \int \frac{1}{B\sqrt{2\pi}} e^{\frac{-\delta^2}{2B^2}} \mathrm{d}\delta.$$

Hence, setting $B = \sqrt{1/(2At)}$ ($A = \eta p^*/2$), via the fact that $1/(1-e^{-x}) \leq 1 + 1/x$ for all $x > 0$, we obtain

$$\begin{aligned}
\mathbb{E}[\Delta^{\boldsymbol{\nu}}_{t+1}] &\leq t^{-\theta} + (3+e^2)m\sqrt{\frac{\pi}{At}} + \int_{t^{-\theta}}^{1} \Big(\frac{(3+e^2)m}{t^2(1-e^{-\delta^2/2})}\Big)\mathrm{d}\delta \\
&\leq t^{-\theta} + (3+e^2)m\sqrt{\frac{\pi}{At}} + \frac{(3+e^2)m}{t^2} \int_{t^{-\theta}}^{1} \Big(1 + \frac{2}{\delta^2}\Big)\mathrm{d}\delta \\
&= t^{-\theta} + (3+e^2)m\sqrt{\frac{\pi}{At}} + (3+e^2)mt^{-2}\Big(\delta - \frac{2}{\delta}\Big)\Big|_{\delta=t^{-\theta}}^{\delta=1} \\
&= t^{-\theta} + (3+e^2)m\sqrt{\frac{\pi}{At}} \\
&+ (3+e^2)mt^{-2}\Big(1 - 2 - t^{-\theta} + 2t^{\theta}\Big) \\
&\leq t^{-\theta} + (3+e^2)m\sqrt{\frac{\pi}{At}} + 2(3+e^2)mt^{\theta-2}.
\end{aligned}$$

The order of t in this bound is minimized when $\theta = 1$; thus, we have

$$\begin{aligned}
\mathbb{E}[\Delta^{\boldsymbol{\nu}}_{t+1}] &\leq (3+e^2)m\sqrt{\frac{\pi}{At}} + (2(3+e^2)m+1)t^{-1} \qquad (16) \\
&\leq (3+e^2)mt^{-1/2}\sqrt{\frac{\pi}{A}} + 3(3+e^2)mt^{-1/2} \\
&\leq (3+e^2)mt^{-1/2}\Big(\sqrt{\frac{\pi}{A}} + 3\Big). \qquad (17)
\end{aligned}$$

□

Next, using Lemma 9, we show that the gap-independent regret for CTS is at most $O(T^{1-w/2})$ for a large set of problems where the bounded-smoothness function $f(x)$ is γx^{ω}, for $\omega \in (0, 1]$.

Note that f is concave. Hence, via Lemma 8, (17) and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{IR}^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(t+1)] &\leq 2\mathbb{E}[f(\Delta^{\boldsymbol{\nu}}_{t+1})] \\
&\leq 2f(\mathbb{E}[\Delta^{\boldsymbol{\nu}}_{t+1}]) \\
&\leq 2f\Big((3+e^2)m\sqrt{\frac{\pi}{At}} + 3(3+e^2)m\sqrt{\frac{1}{t}}\Big) \\
&\leq f\Big(2(3+e^2)m\sqrt{\frac{\pi}{At}}\Big) + f\Big(6(3+e^2)m\sqrt{\frac{1}{t}}\Big) \\
&= \gamma\Big(\big(2(3+e^2)m\big)^{\omega}\Big(\frac{\pi}{At}\Big)^{\omega/2} + \big(6(3+e^2)m\big)^{\omega}t^{-\omega/2}\Big).
\end{aligned}$$

Now, we can bound the regret by bounding the sum of expected instantaneous regrets throughout the time horizon:

$$\begin{aligned}
Reg^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(T) &= \sum_{t=1}^{T} \mathbb{E}[\mathrm{IR}^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(t)] \\
&\leq \sum_{t=1}^{\lceil t' \rceil} \mathbb{E}[\mathrm{IR}^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(t)] + \sum_{t=\lceil t' \rceil+1}^{T} \mathbb{E}[\mathrm{IR}^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(t)] \\
&\leq \lceil t' \rceil \nabla_{\max} + \gamma\Big(2m(3+e^2)\Big)^{\omega} \left[\left(\frac{\pi}{A}\right)^{\omega/2} + 3^{\omega} \right] \sum_{t=1}^{T} t^{-\omega/2}.
\end{aligned}$$

Using the fact that $\sum_{t=1}^{T} t^{-x} \leq T^{1-x}/(1-x)\ \forall x \in (0,1)$, we have the following for $\omega \in (0,1]$:

$$Reg^{\mathrm{CTS}}_{\boldsymbol{\mu},\alpha,\beta}(T) \leq \lceil t' \rceil \nabla_{\max} + \gamma\Big(2m(3+e^2)\Big)^{\omega} \left[\left(\frac{\pi}{A}\right)^{\omega/2} + 3^{\omega} \right] \frac{T^{1-w/2}}{1-w/2}.$$

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. In: *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp 2312–2320

Agrawal R (1995) Sample mean based index policies by $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability* 27(4):1054–1078

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. In: *Proc. Conference on Learning Theory (COLT)*, pp 39.1–39.26

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2017) Thompson sampling for the MNL-bandit. *arXiv preprint arXiv:170600977*

Akbarzadeh N, Tekin C (2016) Gambler's ruin bandit problem. In: *Proc. 54th Annual Allerton Conference on Communication, Control, and Computing*, pp 1236–1243

Anantharam V, Varaiya P, Walrand J (1987) Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards. *IEEE Trans Autom Control* 32(11):968–976

Atan O, Tekin C, Schaar Mvd (2015) Global multi-armed bandits with Hölder continuity. In: *Proc. AISTATS*, pp 28–36

Auer P (2003) Using confidence bounds for exploitation-exploration trade-offs. *J Mach Learn Res (JMLR)* 3:397–422

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2-3):235–256

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5(1):1–122

Chen W, Wang Y, Yuan Y (2013) Combinatorial multi-armed bandit: General framework and applications. In: *Proc. Int. Conf. Machine Learning*, pp 151–159

Chen W, Hu W, Li F, Li J, Liu Y, Lu P (2016a) Combinatorial multi-armed bandit with general reward functions. In: *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp 1651–1659

Chen W, Wang Y, Yuan Y, Wang Q (2016b) Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *J Mach Learn Res (JMLR)* 17(1):1746–1778

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. In: *Proc. Conference on Learning Theory (COLT)*, pp 355–366

Durand A, Gagné C (2014) Thompson sampling for combinatorial bandits and its application to online feature selection. In: *Proc. AAAI-14 Workshop on Sequential Decision-making with Big Data*

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

Gopalan A, Mannor S, Mansour Y (2014) Thompson sampling for complex online problems. In: *Proc. 31st Int. Conf. Machine Learning (ICML)*, vol 32, pp 100–108

Graepel T, Quiñonero Candela J, Borchert T, Herbrich R (2010) Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing Search Engine. In: *Proc. 27th Int. Conf. Machine Learning (ICML)*, pp 13–20

Granmo O (2010) Solving two-armed Bernoulli bandit problems using a Bayesian learning automaton. *Int J Intelligent Computing and Cybernetics* 3(2):207–234

Kalman D (2001) A generalized logarithm for exponential-linear equations. *The College Mathematics Journal* 32(1)

Kaufmann E, Korda N, Munos R (2012) Thompson sampling: An optimal finite time analysis. In: *Proc. Int. Conf. Algorithmic Learning Theory (ALT)*, pp 199–213

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: *Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp 137–146

Kveton B, Wen Z, Ashkan A, Eydgahi H, Eriksson B (2014) Matroid bandits: Fast combinatorial optimization with learning. *arXiv preprint arXiv:14035045*

Kveton B, Szepesvari C, Wen Z, Ashkan A (2015a) Cascading bandits: Learning to rank in the cascade model. In: *Proc. 32nd Int. Conf. Machine Learning (ICML)*, pp 767–776

Kveton B, Wen Z, Ashkan A, Szepesvari C (2015b) Combinatorial cascading bandits. In: *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp 1450–1458

Kveton B, Wen Z, Ashkan A, Szepesvári C (2015c) Tight regret bounds for stochastic combinatorial semi-bandits. In: *Proc. AISTATS*, vol 38, pp 535–543

Lai T, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4–22

Lei S, Maniu S, Mo L, Cheng R, Senellart P (2015) Online influence maximization. In: *Proc. 21th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp 645–654

Mersereau AJ, Rusmevichientong P, Tsitsiklis JN (2009) A structured multiarmed bandit problem and the greedy policy. *IEEE Trans Autom Control* 54(12):2787–2802

Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions–I. *Mathematical Programming* 14(1):265–294

Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535

Sarıtaç A, Karakurt A, Tekin C (2016) Online contextual influence maximization in social networks. In: *Proc. 54th Annual Allerton Conference on Communication, Control, and Computing*, pp 1204–1211

Scott SL (2010) A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26(6):639–658

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294

Vaswani S, Lakshmanan L, Schmidt M, et al (2015) Influence maximization with bandits. *arXiv preprint arXiv:150300024*

Vazirani VV (2001) *Approximation algorithms*. Springer

Wang Q, Chen W (2017) Tighter regret bounds for influence maximization and other combinatorial semi-bandits with probabilistically triggered arms. *arXiv preprint arXiv:170301610*

Wen Z, Kveton B, Valko M (2017) Online influence maximization under independent cascade model with semi-bandit feedback. *arXiv preprint arXiv:160506593*