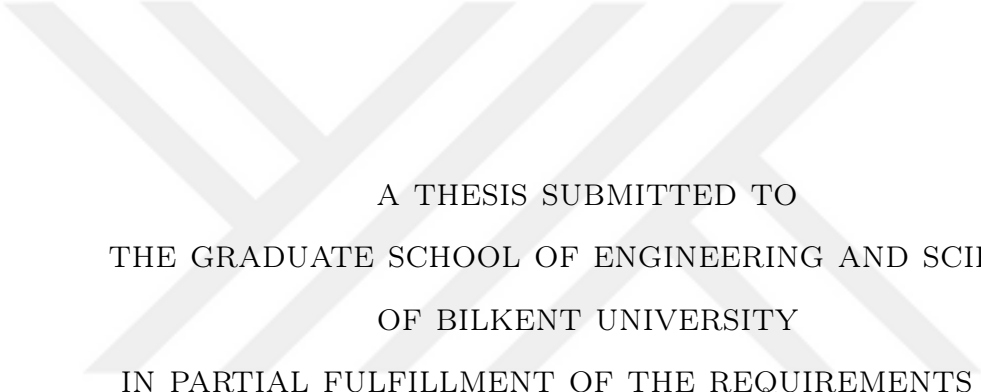


COMBINATORIAL MULTI-ARMED BANDITS: APPLICATIONS AND ANALYSES



A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

By
Anıl Ömer Sarıtaç
September 2018

Combinatorial Multi-armed Bandits: Applications and Analyses

By Anıl Ömer Sarıtaç

September 2018

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Savaş Dayanık(Advisor)

Cem Tekin(Co-Advisor)

Ülkü Gürler

Elif Vural

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

COMBINATORIAL MULTI-ARMED BANDITS: APPLICATIONS AND ANALYSES

Anıl Ömer Sarıtaç
M.S. in Industrial Engineering
Advisor: Savaş Dayanık
Co-Advisor: Cem Tekin
September 2018

We focus on two related problems: *Combinatorial multi-armed bandit problem (CMAB) with probabilistically triggered arms (PTAs)* and *Online Contextual Influence Maximization Problem with Costly Observations (OCIMP-CO)* where we utilize a CMAB approach. 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 gap-dependent and $O(\sqrt{T})$ gap-independent regret improving on previous works which study CMAB with PTAs under more general ATPs. Then, we numerically evaluate the performance of CUCB- κ and CTS in a real-world movie recommendation problem. For the *Online Contextual Influence Maximization Problem with Costly Observations*, we study a case where the learner can observe the spread of influence by paying an *observation cost*, by which it aims to maximize the total number of influenced nodes over all epochs minus the observation costs. Since the offline influence maximization problem is NP-hard, we develop a CMAB approach that use an approximation algorithm as a subroutine to obtain the set of seed nodes in each epoch. When the influence probabilities are Hölder continuous functions of the context, we prove that these algorithms achieve sublinear regret (for any sequence of contexts) with respect to an approximation oracle that knows the influence probabilities for all contexts. Moreover, we prove a lower bound that matches the upper bound with respect to time and cost order, suggesting that the upper bound is the best possible. Our numerical results on several networks illustrate that the proposed algorithms perform on par with the state-of-the-art methods even when the observations are cost-free.

Keywords: Combinatorial Bandits, Multi-armed Bandit, Approximation Algorithms, Probabilistically Triggered Arms, Influence Maximization, Costly Observations, Regret Bounds, Lower Bound.



ÖZET

KOMBINATORİK ÇOK KOLLU HAYDUTLAR: UYGULAMALAR VE ANALİZLER

Anıl Ömer Sarıtaç
Endüstri Mühendisliği, Yüksek Lisans
Tez Danışmanı: Savaş Dayanık
İkinci Tez Danışmanı: Cem Tekin
Eylül 2018

Birbiriyle yakından ilişkili iki probleme odaklanıyoruz: Kolları olasılıksal olarak tetiklenen Kombinatorik Çok Kollu Haydut Problemi (KÇKHP) ve bir KÇKHP yaklaşımı getirdiğimiz Kontekst içeren Gerçek zamanlı Etki Maksimizasyonu Problemi (KGEMP). Kolları tetikleme olasılıklarının (KTO) pozitif olduğunu varsaydığımız KÇKHP probleminde, bu konularda sıklıkla çalışılan bir grup algoritmanın performans garantilerini bularak literatürdeki, KTOlar daha zayıf varsayımlar altındayken bulunan performans garantilerini geliştiriyoruz. Sonrasında ise bir gerçek hayat film önerisi sisteminde bu algoritmaları numerik olarak değerlendiriyoruz. KGEMP problemi için ise, öğrencinin etki yayılımını bir bedel ödeyerek öğrendiği ve bu yolla, bütün dönemlerdeki toplam etkilenen düğüm sayısından gözlem bedel değeri çıkarılarak elde edilen amaç fonksiyonunu maksimize etmeye çalışan bir problem durumunu çalışıyoruz. Gerçek zamanlı olmayan Etki Maksimizasyonu problemi NP-Zor olduğundan, bir yaklaşım algoritmasını alt-rutin olarak kullanan bir KÇKHP problemi öneriyoruz. Etkileme olasılıklarının kontekstin Hölder-süreklili bir fonksiyonu olduğu durumda, pişmanlığın zamanın lineer bir fonksiyonunun altında kalacağını ispatlıyoruz. Dahası, pişmanlık için bir alt sınır bularak bulduğumuz üst sınırın elde edilebilecek en iyi üst sınır olduğunu kanıtlıyoruz. Numerik sonuçlarımız ile ise önerdiğimiz algoritmaların, gözlemlerin bedelsiz olduğu durumlarda bile literatürdeki en iyi algoritmalar ile başabaş bir performans gösterdiğini görüyoruz.

Anahtar sözcükler: Kombinatorik Haydutlar, Çok kollu haydut, Yaklaşım algoritmaları, Olasılıksal olarak tetiklenen kollar, Etki Maksimizasyonu, Gözlem bedelleri, Pişmanlık sınırları, alt sınırlar.

Acknowledgement

I am sincerely grateful for the support of my advisors Prof. Savaş Dayanık and Asst. Prof. Cem Tekin, which pushed me to do quality work.

In addition, I would like to thank Prof. Ülkü Gürler and Asst. Prof. Elif Vural for their valuable participations in the jury.

...

Contents

1	Introduction	1
1.1	Introduction to Online Contextual Influence Maximization Problem with Costly Observations (OCIMP-CO)	2
1.2	Introduction to Combinatorial Multi-armed Bandit (CMAB) problem	6
2	Literature Review	10
2.1	Influence Maximization (IM)	10
2.2	Multi-armed Bandit (MAB) and Applications to Influence Maximization (IM)	11
3	Problem Formulations	16
3.1	CMAB with PTAs	16
3.2	Online Contextual Influence Maximization with Costly Observations(OCIMP-CO)	18
3.2.1	Definition of the Influence	18
3.2.2	Definition of the Reward and the Regret	19

4	Algorithms	22
4.1	Algorithms for CMAB with PTAs	22
4.1.1	CUCB- κ	22
4.1.2	CTS	24
4.2	Algorithms for Online Contextual Influence Maximization Problem	25
4.2.1	Contextual Online Influence Maximization with Costly Edge-level Feedback (COIN-CO-EL)	26
4.2.2	Contextual Online Influence Maximization with Costly Node-level Feedback (COIN-CO-NL)	29
5	Regret Analysis	32
5.1	Regret Analysis of Algorithms for CMAB with PTAs	32
5.1.1	Regret Analysis for CUCB- κ	33
5.1.2	Regret Analysis for CTS	36
5.2	Regret Analysis of Algorithms for Online Contextual Influence Maximization Problem with Costly Observations(OCIMP-CO) . .	37
5.2.1	Approximation Guarantee	38
5.2.2	Regret analysis for COIN-CO-EL	38
5.2.3	Regret analysis for COIN-CO-NL	40
5.2.4	Lower Bound on the Regret for the Context-free Online IM Problem	41

6	Experimental Results	43
6.1	Online Contextual Influence Maximization Problem with Costly Observations(OCIMP-CO)	43
6.1.1	Extensions	43
6.1.2	Experiments	45
6.1.3	Results for Cost-Free Edge-level Feedback	51
6.1.4	Results for Cost-Free Node-level Feedback	52
6.1.5	Results for the Costly Edge-level Feedback	53
6.2	CMAB with PTAs	54
6.2.1	Definition of the Recommendation Problem	54
6.2.2	Calculation of the Influence Probabilities	55
6.2.3	Results	56
7	Conclusion	58
A	Appendix	67
A.1	Proofs of Theorems Regarding Algorithms for the CMAB with PTAs	67
A.1.1	Preliminaries	67
A.1.2	Proof of Theorem 1	68
A.1.3	Proof of Theorem 2	70
A.1.4	Proof of Theorem 3	74

A.1.5	Proof of Theorem 4	77
A.1.6	Proof of Theorem 5	80
A.1.7	Proof of Theorem 6	83
A.1.8	Proof of Theorem 7	85
A.2	Proofs of Theorems Regarding Algorithms for OCIMP	88
A.2.1	Proof Of Theorem 9	88
A.2.2	Proof of Theorems 10 and 11	89
A.2.3	Proof of Theorem 12	97

Published Articles

The work presented in Sections 1.1, 2.1, 3.2, 4.2, 5.2, 6.1 and Appendix A.2 of this thesis has appeared in publications (1) and (3) listed below whereas the rest of the thesis has appeared in (2).

1. © 2016 IEEE. Reprinted, with permission, from A. O. Saritac, A. Karakurt, and C. Tekin, “Online contextual influence maximization in social networks,” in Proc. 54th Annual Allerton Conf. Communication, Control, and Computing (Allerton), pp. 1204–1211, 2016.
2. © 2017 IEEE. Reprinted, with permission, from A. O. Saritac and C. Tekin, “Combinatorial multi-armed bandit problem with probabilistically triggered arms: A case with bounded regret,” in 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 111–115, 2017.
3. © 2018 IEEE. Reprinted, with permission, from A. O. Saritac, A. Karakurt, and C. Tekin, “Online contextual influence maximization with costly observations,” IEEE Transactions on Signal and Information Processing over Networks, 2018. DOI:10.1109/TSIPN.2018.2866334.

List of Figures

1.1	An illustration that shows the influence spread process for $k = 1$ and time slots $s = 1, 2, 3$ in epoch t . Numbers on the edges denote the influence probabilities. For this example, the influence spread in epoch t is 2, and the expected influence spread is $0.1 + 0.1 + 0.9 + 0.9 * 0.9 = 1.91$. $\mathcal{R}_t^s$ denotes the set of nodes influenced in time slot s of epoch t . $\mathcal{R}_t^1$ denotes the seed node. $\mathcal{A}_t^s$ denotes the set of nodes influenced prior to time slot s of epoch t . $\mathcal{C}_t^{s+1}$ denotes the set of nodes that might be influenced in time slot $s + 1$ of epoch t .	5
6.1	Results for context-aware algorithms under cost-free edge-level feedback (a-d), cost-free node-level feedback (e-f) and costly edge-level feedback (g-h).	48
6.2	Average regrets of COIN-CO-EL+ and COIN-CO-EL-HD calculated over exploration and exploitation epochs.	53
6.3	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).	54

List of Tables

1.1	Table of Notations for OCIMP-CO	3
1.2	Comparison of our work on OCIMP-CO with prior works.	5
1.3	Table of Notations for CMAB with PTAs	7
1.4	Comparison of our work on CMAB-PTAs with prior works.	9
2.1	Comparison of our work on CMAB-PTAs with prior works.	13
2.2	Comparison of our work on OCIMP-CO with prior works.	14
6.1	Properties of the networks used in the experiments.	47

Chapter 1

Introduction

The purpose in this thesis is to devise effective learning algorithms (learners) that have theoretical performance guarantees as well as competitive experimental performance for problems where the environment (parameters of the problem) is not known in the beginning. These learners utilize approximation algorithms that are designed for solving the problems for the case where the environment is known. Through continuously taking an action and observing the reaction in the environment, they learn the environment; therefore, take better actions as time passes. In Section 1.2, we give an example of this process via an application of CMAB with PTAs for the Online Influence Maximization problem, where the learner chooses a set of nodes in each epoch, and updates its choice as it gathers more information about the process of influence spread.

We introduce two connected problems for this purpose. In Section 1.1, we introduce the OCIMP-CO problem whereas in Section 1.2, we introduce the CMAB with PTAs problem. As explained explicitly in Section 1.2, CMAB with PTAs is closely related with the OCIMP-CO problem, where the cost of observations and context is not considered. Importantly, this problem is not the only application for CMAB with PTAs as it has applications in wireless networking, online advertising, and recommendation problems [1–7]. Therefore, studying CMAB-PTAs allows us to conclude that the theoretical guarantees we find are not only valid for the

Influence Maximization problem but also valid for other applications of CMAB with PTAs.

1.1 Introduction to Online Contextual Influence Maximization Problem with Costly Observations (OCIMP-CO)

In recent years, there has been growing interest in understanding how influence spreads in a social network [8–14]. This interest is motivated by the proliferation of viral marketing in social networks. For instance, nowadays many companies promote their products on social networks by giving free samples of certain products to a set of seed nodes/users, expecting them to influence people in their social circles into purchasing these products. The objective of these companies is to find out the set of nodes that can collectively influence the greatest number of other nodes in the social network. This problem is called the *influence maximization* (IM) problem.

In the IM problem, the spread of influence is modeled by an *influence graph*, where directed edges between nodes represent the paths that the influence can propagate through and the weights on the directed edges represent the likelihood of the influence, i.e., the influence probability. Numerous models are proposed for the spread of influence, with the most popular ones being *independent cascade* (IC) and *linear threshold* (LT) models [15]. In the IC model, the influence propagates on each edge independently from the other edges of the network, and an influenced node has only a single chance to influence its neighbors. Hence, only recently influenced nodes can propagate the influence. Thus, the influence stops to spread when the recently influenced nodes fail to influence their neighbors. On the other hand, in the LT model, a node’s chance to get influenced depends on whether the sum of weights of its active neighbors exceeds a threshold or not.

Most of the prior work in IM assume that the influence probabilities of the

Notation	Description
$G(V, E)$	Graph representing the social network.
V	Set of nodes in $G(V, E)$.
E	Set of edges in $G(V, E)$.
n	Number of nodes in V .
m	Number of edges in E .
$\mathcal{X}$	Context set with support $[0, 1]^d$, which is a metric space.
d	Number of dimensions for any $x \in \mathcal{X}$
$p_{i,j}^x$	Influence probability for the directed edge $(i, j) \in E$ under context x .
$\sigma(x, S)$	Expected number of influenced nodes for context $x \in \mathcal{X}$ and seed set S .
$S^*(x)$	The optimal seed set(action) under context $x \in \mathcal{X}$.
c	Fixed cost incurred for an observation.
B_t	Number of observations made in epoch t .
$R_\pi^{(\alpha, \beta)}(T)$	Regret for learner π by epoch T when it uses an (α, β) -approx. algorithm.
$Rew^*(T)$	Expected optimal cumulative reward by epoch T .

Table 1.1: Table of Notations for OCIMP-CO

influence graph are known and that the influence spread process is observed [16–21], and focus on designing computationally efficient algorithms to maximize the influence spread. However, in many practical settings, it is impossible to know beforehand the influence probabilities exactly. For instance, a firm that wants to introduce a new product or to advertise its existing products in a new social network may not know the influence probabilities on the edges of the network. In contrast to the prior works mentioned above, our focus is to design an optimal learning strategy when the influence probabilities are unknown.

In the marketing example given above, influence depends on the product that is being advertised as well as the identities of the users. Hence, the characteristic (context) of the product affects the influence probabilities. The strand of literature that is closest to the problem we consider in this thesis in terms of the dependence of the influence probabilities on the context is called *topic-aware IM* [10–13]. To the best of our knowledge, none of the prior works in topic-aware IM develop learning algorithms with provable performance guarantees for the case when the influence probabilities are unknown. In addition, prior works in IM that consider unknown influence probabilities do not consider the cost of feedback (cost of observation of the influence spread process) [22–26]. However, this cost

exists in most of the real-world applications of IM. For instance, finding out who influenced a specific person into buying a product might require conducting a costly investigation (e.g., a survey).

Motivated by such real-world applications, in this thesis, we define a new learning model for IM, called the *Online Contextual Influence Maximization Problem with Costly Observations* (OCIMP-CO). In contrast to IM, which is a single-shot problem, OCIMP-CO is a sequential decision making problem. In OCIMP-CO, the learner (e.g., the firm in the above example), faces a series of epochs in each of which a different influence campaign is run. At the beginning of each epoch, the learner observes the context of that epoch. For instance, the context can be the type of the influence campaign (e.g., one influence campaign might promote a sports equipment, while another influence campaign might promote a mobile data plan). After observing the context, the learner chooses a set of k seed nodes to influence. We call these nodes *exogenously influenced nodes*. Then, the influence spreads according to the IC model, which is explained in detail in Section 3.2.1. The nodes that are influenced as a result of this process are called *endogenously influenced nodes*. An illustration of the influence spread process is given in Fig. 1.1. At the end of each epoch, the learner obtains as its reward the number of endogenously influenced nodes. The goal of the learner is to maximize its cumulative expected influence spread minus the observation costs over epochs.

In this thesis, we consider two different influence observation settings: costly edge-level feedback, in which the learner freely observes the set of influenced nodes, but pays to observe the influence outcomes on the edges of the network; and costly node-level feedback, in which the learner pays to observe whether a node is influenced or not. For the costly edge-level feedback setting we propose a learning algorithm called *Contextual Online INfluence maximization COstly Edge-Level feedback* (COIN-CO-EL) to maximize the learner’s reward for any given number of epochs. COIN-CO-EL can use any approximation algorithm for the offline IM problem as a subroutine to obtain the set of seed nodes in each epoch. When the influence probabilities are Hölder continuous functions of the context, we prove that COIN-CO-EL achieves $O(c^{1/3}T^{(2\theta+d)/(3\theta+d)})$ regret (for any sequence

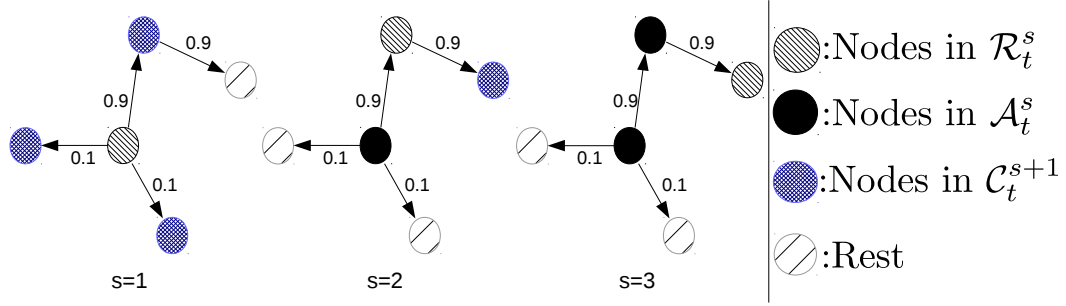


Figure 1.1: An illustration that shows the influence spread process for $k = 1$ and time slots $s = 1, 2, 3$ in epoch t . Numbers on the edges denote the influence probabilities. For this example, the influence spread in epoch t is 2, and the expected influence spread is $0.1 + 0.1 + 0.9 + 0.9 \times 0.9 = 1.91$. $\mathcal{R}_t^s$ denotes the set of nodes influenced in time slot s of epoch t . $\mathcal{R}_t^1$ denotes the seed node. $\mathcal{A}_t^s$ denotes the set of nodes influenced prior to time slot s of epoch t . $\mathcal{C}_t^{s+1}$ denotes the set of nodes that might be influenced in time slot $s + 1$ of epoch t .

of contexts) with respect to an approximation oracle that knows the influence probabilities for all contexts. Here, c represents the observation cost, θ is the exponent of Hölder condition for the influence probabilities, and d represents the dimension of the context. Then, we also propose an algorithm for the costly node-level feedback setting, called *Contextual Online INfluence maximization COstly Node-Level feedback* (COIN-CO-NL), which learns the influence probabilities by performing smart explorations over the influence graph. We also show that COIN-CO-NL enjoys $O(c^{1/3}T^{(2\theta+d)/(3\theta+d)})$ regret. In addition, we prove that for the special case when the influence probabilities do not depend on the context, i.e., the context-free online IM problem with costly observations, our algorithms achieve $O(c^{1/3}T^{2/3})$ regret. We conclude that this bound is tight in terms of the observation cost and the time order by proving that the regret lower bound for this case is $\Omega(c^{1/3}T^{2/3})$. Table 1.2 gives a comparison of our results for OCIMP-CO with the results in the literature.

Table 1.2: Comparison of our work on OCIMP-CO with prior works.

	Our Work	[24, 25, 27]	[10–13]	[15–19, 21, 28–30]	[31]
Context	Yes	No	Yes	No	No
Online Learning	Yes	Yes	No	No	Yes
Regret Bound	Yes	Yes	No	No	No
Costly Observation	Yes	No	No	No	No

1.2 Introduction to Combinatorial Multi-armed Bandit (CMAB) problem

The 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 past [32–36]. 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. Combinatorial multi-armed bandit (CMAB) framework is suitable for these applications, where 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 usually observes the states of the chosen arms. In the semi-bandit feedback case, the learner only partially observes the states of the chosen arms. The applications for CMAB include wireless networking, online advertising and recommendation, and viral marketing. [1–7].

A rather popular application for CMAB is Cascading Bandits [6]. This problem models the click behaviour of a user during web search. In the model, the user examines a list of recommended items (such as web pages) from the first item to the last, and selects the first attractive item. Therefore, the items the user observe before the selection are unattractive, and the items after it remains unobserved. Consequently, the aim of a recommendation engine that selects the list of K items

Notation	Description
$X_i^{(t)}$	State of arm i at epoch t .
$\mathbf{X}^{(t)}$	Vector comprised of $X_i^{(t)}$'s.
S_t	Set of arms chosen in epoch t by the learner as the action set.
τ_t	Random set of triggered arms, which is drawn from $D^{\text{trig}}(S_t, \mathbf{X}^{(t)})$.
p^*	Minimum triggering probability : $\Pr(i \in \tau_t) \geq p^* \forall i \in \{1, \dots, m\}$.
m	Number of arms.
$R(S_t, \mathbf{X}^{(t)}, \tau_t)$	Finite non-negative reward, depends deterministically on S_t , $\mathbf{X}^{(t)}$ and τ_t .
μ_i	Expected state of arm i : $\mu_i := \mathbb{E}_{\mathbf{X}^{(t)} \sim D}[X_i^{(t)}]$.
$\boldsymbol{\mu}$	Expectation vector : $\boldsymbol{\mu} := (\mu_1, \dots, \mu_m)$.
$r_{\boldsymbol{\mu}}(S)$	Expected reward function : $r_{\boldsymbol{\mu}}(S) := \mathbb{E}[R(S, \mathbf{X}, \tau)]$.
S^*	An optimal action : $S^* \in \arg \max_{S \in \mathcal{S}} r_{\boldsymbol{\mu}}(S)$.
$r_{\boldsymbol{\mu}}^*$	Expected reward of the optimal action.
$\text{Reg}_{\boldsymbol{\mu}, \alpha, \beta}^{\pi}(T)$	Regret of algorithm π that uses an (α, β) -approx. algorithm.

Table 1.3: Table of Notations for CMAB with PTAs

is to maximize the probability that the user finds at least one attractive item in the list. In the CMAB framework, chosen arms correspond to the list of items shown to the user, and the reward is a non-linear function of the states of the chosen arms. This problem is more challenging than the classical MAB problem due to the fact that the size of the action set is combinatorial in the number of arms, and the reward is possibly a non-linear function of the states of the chosen arms.

An interesting extension to the CMAB is CMAB with PTAs [24, 37], 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 [38], 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.

An application of CMAB with PTAs is the Online Influence Maximization (OIM) problem. Differently from OCIMP-CO, which is introduced in the previous

chapter, OIM does not consider cost of observations and the context information. It is primarily motivated by the growing interest in viral marketing, where a sample of the product to be advertised are given to people with the highest potential to influence their social circles into buying the product. Motivated by this, the problem of finding these most influential people comprises the goal of the OIM problem. In OIM, we are given a graph $G(V, E)$ which represents a social network, where the set of nodes corresponds to the people in the network whereas the set of edges corresponds to the connections between them. The aim is to choose a set (action set) S with fixed cardinality that maximizes the number of influenced nodes (people) in the network. Whether a node is influenced as a result of the action S is determined by the probabilities defined for each edge in the network. Hence, depending on the set S , each node gets influenced with some probability. This problem is an instance of CMAB with PTAs. In the CMAB with PTAs, the action set, which is a set of arms, corresponds to the people who take the sample product to be advertised. Moreover, the triggering probability for an arm given an action set corresponds to probability of the event that a node becomes influenced as a result of the choice of the action set.

In this thesis, we consider an instance of the 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 the posterior distribution based on 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 [39, 40]. However, these works do not conflict with the asymptotic $O(\log T)$ lower bound for the classical MAB problem [34], 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. In Table 1.4, we compare our theoretical results with the theoretical results in the literature.

Table 1.4: Comparison of our work on CMAB-PTAs with prior works.

	Our Work	[24, 38]	[5, 7, 41, 42]	[4, 6]
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

Chapter 2

Literature Review

In OCIMP-CO, an Influence Maximization problem is solved in each epoch. In addition, both OCIMP-CO and CMAB with PTAs are frameworks where Multi-armed Bandit approaches are utilised. Hence, our work is related with the literature on the Influence Maximization problem as well as the Multi-armed Bandit problems. We first discuss the literature on the Influence Maximization problem, then discuss the Multi-armed Bandit problem and applications of it for the Influence Maximization problem.

2.1 Influence Maximization (IM)

The IM problem was first proposed in [15], where it is proven to be NP-Hard and an approximately optimal solution is given. However, the solution given in [15] does not scale well because it often requires thousands of Monte Carlo samples to estimate the expected influence spread of each seed set. This motivated the development of many heuristic methods with lower computational complexity [17, 21, 28, 29].

In numerous other works, algorithms with approximation guarantees are developed for the IM problem, such as CELF [18], CELF++ [19] and NewGreedy [21].

In addition to these works, in [30], an approximation algorithm based on reverse influence sampling is proposed and its run-time optimality is proven. In [16], the authors improved the scalability of this algorithm by proposing two new algorithms TIM and TIM+. More recently, [20] developed IMM which is an improvement on TIM in terms of efficiency while preserving its theoretical guarantees. None of the works mentioned above consider the context information. IM based on context information is studied in several other works such as [10, 12, 13]. However, in contrast to our work which solves a more general problem, these works assume that the influence probabilities are known and topics/contexts are discrete. Moreover, in OCIMP-CO, context is represented by a collection of continuous features (which can be discretized if necessary). It is also worth to mention that, to the best of our knowledge, there exists no work that solves the online version of the IM problem where observing the influence spread process is costly.

2.2 Multi-armed Bandit (MAB) and Applications to Influence Maximization (IM)

Two key techniques are used for learning in the MAB problem: UCB-based index policies [34] and Thompson (posterior) sampling [32]. [34] introduced UCB-based index policies for the MAB problem and proved a tight logarithmic bound on the asymptotic regret, which establishes asymptotic optimality of the proposed set of policies. Later on, [43] revealed that sample-mean based index policies can also achieve $O(\log T)$ regret, and [35] 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 [44] and [45], 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 [46–48], 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 [5, 41, 49–51]. In addition, [6] and [7] solve specific instances of CMAB problems with nonlinear reward functions. Moreover, similar to our work, [24] and [38] consider the 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, [4] 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 [52], 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 [53]. 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 2.1.

Influence maximization (IM) problem is closely related to CMAB [15]. Various works consider the online version of this problem, the OIM problem [22, 23, 31]. In this version, the ATPs are unknown a priori, and as in ours, the set of arms

chosen at each epoch corresponds to the seed set of nodes. Works such as [24] and [38] solve this problem by using algorithms developed for the CMAB with PTAs. Differently from these, [22] 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.

Table 2.1: Comparison of our work on CMAB-PTAs with prior works.

	Our Work	[24, 38]	[5, 7, 41, 42]	[4, 6]
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

In [24], the authors present a combinatorial MAB problem where multiple arms are chosen at each epoch, and these arms probabilistically trigger the other arms. In our terminology, multiple arms chosen at each epoch correspond to the set of seed nodes and probabilistically triggered arms correspond to nodes other than the set of seed nodes. For this problem, a logarithmic gap-dependent regret bound is proven with respect to an approximation oracle. In a subsequent work, the dependence of the regret on the inverse of the minimum positive arm triggering probability is removed under more stringent assumptions on the reward function [54]. However, the problem in [24] does not involve any contexts.

Another general MAB model that uses greedy algorithms to solve the IM problem with unknown graph structure and influence probabilities is proposed in [25]. In addition, [27] considers a non-stationary IM problem, in which the influence probabilities are unknown and time varying. OCIMP-CO is more general than this, since the context can also be used to model the time-varying nature of the influence probabilities (for instance, one dimension of the context can be the time).

An online method for the IM problem that uses an *upper confidence bound* (UCB) based and an ϵ -greedy based algorithm is proposed in [31], but theoretical analysis of this method is not carried out. In another related work [22], the IM problem is defined on an undirected graph where the influence probabilities are assumed to be linear functions of the unknown parameters, and a linear

UCB-based algorithm is proposed to solve it. The prior works described above assume that the influence outcomes on each edge in the network are observed by the learner. Recently, another observation model, called node-level feedback, is proposed in [23]. This model assumes that only the influenced nodes are observable while the spread of influence over the edges is not. However, no regret analysis is provided for this model.

There also exists another strand of literature that studies contextual MAB and its combinatorial variants under the linear realizability assumption [55–57]. This assumption enforces the relation between the expected rewards (also known as scores in combinatorial MAB literature) of the arms and the contexts to take a linear form, which boils down learning to estimating an unknown parameter vector. This enables the development of learning algorithms that can achieve $\tilde{O}(\sqrt{T})$ regret. While [55] directly models the expected reward of an arm as a linear function of the context, [56] and [57] consider the combinatorial MAB problem where the expected reward of an action is a monotone and Lipschitz continuous function of the expected scores of the arms associated with the action. This model is more restrictive than ours since it forces the arm scores (i.e., the influence probabilities in our setting) to be linear in the context. In contrast, in our work, we only assume that the influence probabilities are Hölder continuous functions of the context (see Assumption 3).

Table 2.2: Comparison of our work on OCIMP-CO with prior works.

	Our Work	[24, 25, 27]	[10–13]	[15–19, 21, 28–30]	[31]
Context	Yes	No	Yes	No	No
Online Learning	Yes	Yes	No	No	Yes
Regret Bound	Yes	Yes	No	No	No
Costly Observation	Yes	No	No	No	No

In conclusion, our work on OCIMP-CO differentiates itself by considering context as well as the cost of observation in the online IM problem. The differences between our work and the prior works are summarized in Table 2.2. Moreover, the theoretical results we prove for CMAB in this thesis also applies to the OIM problem defined over a strongly connected graph where each node is reachable

from the other nodes in the graph.



Chapter 3

Problem Formulations

3.1 CMAB with PTAs

First, we define the CMAB with PTAs. We adopt the notation in [38]. The system operates in discrete epochs indexed by t . There are m arms, given by the set $\{1, \dots, m\}$, 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 $\mathbf{X}^{(t)} := (X_1^{(t)}, \dots, X_m^{(t)})$.

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 $\tau_t \subseteq \{1, \dots, m\}$ is triggered based on S_t and $\mathbf{X}^{(t)}$. Here, τ_t is drawn from a multivariate distribution $D^{\text{trig}}(S_t, \mathbf{X}^{(t)})$ called the *probabilistic triggering function* known by the learner (see [38] for examples), where $\Pr(i \in \tau_t) \geq p^*$ for some $p^* > 0$, for all $i \in \{1, \dots, m\}$, 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 $R(S_t, \mathbf{X}^{(t)}, \tau_t)$ that depends deterministically on S_t , $\mathbf{X}^{(t)}$ and τ_t , and observes the states of the triggered arms, i.e., $X_i^{(t)}$, $i \in \tau_t$. R is also known by the learner

(see [38] for examples). The goal of the learner is to maximize its total expected reward over all epochs.

For each arm $i \in \{1, \dots, m\}$, we let $\mu_i := \mathbb{E}_{\mathbf{X}^{(t)} \sim D}[X_i^{(t)}]$ denote the expected state of arm i and $\boldsymbol{\mu} := (\mu_1, \dots, \mu_m)$ denote the *expectation vector*. The expected reward of action S is $r_{\boldsymbol{\mu}}(S) := \mathbb{E}[R(S, \mathbf{X}, \tau)]$ where the expectation is taken over $\mathbf{X} \sim D$ and $\tau \sim D^{\text{trig}}(S, \mathbf{X})$. 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 are known is often an NP-hard problem for which (α, β) -approximation algorithms exist [58]. 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 $\Pr(r_{\boldsymbol{\mu}}(S^{\mathcal{O}}) \geq \alpha r_{\boldsymbol{\mu}}^*) \geq \beta$. 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:

$$\text{Reg}_{\boldsymbol{\mu}, \alpha, \beta}^{\pi}(T) := T\alpha\beta r_{\boldsymbol{\mu}}^* - \mathbb{E} \left[\sum_{t=1}^T r_{\boldsymbol{\mu}}(S_t) \right]. \quad (3.1)$$

For the purpose of regret analysis, as in [24], 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 ([24]). $\exists f : \mathbb{R}^+ \cup \{0\} \rightarrow \mathbb{R}^+ \cup \{0\}$ 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 $\max_{i \in \{1, \dots, m\}} |\mu_i - \mu'_i| \leq \Delta$, $\forall S \in \mathcal{S}$.

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

Assumption 2 ([24]). If for all arms $i \in \{1, \dots, m\}$, $\mu_i \leq \mu'_i$, then we have $r_{\boldsymbol{\mu}}(S) \leq r_{\boldsymbol{\mu}'}(S)$, $\forall S \in \mathcal{S}$.

Next, we define the OCIMP-CO problem.

3.2 Online Contextual Influence Maximization with Costly Observations(OCIMP-CO)

3.2.1 Definition of the Influence

Consider a learner operating on a network with n nodes/users and m edges. The set of nodes is denoted by V and the set of edges is denoted by E . The *network graph* is denoted by $G(V, E)$. The set of children of node i is given by $\mathcal{N}_i := \{j \in V : (i, j) \in E\}$, and the set of parents of node i is given by $\mathcal{V}_i := \{j \in V : (j, i) \in E\}$.

The system operates over time in discrete epochs, indexed by $t \in \{1, 2, \dots\}$. Without loss of generality, context of at t th epoch comes from a d -dimensional context set $\mathcal{X} := [0, 1]^d$, and is denoted by x_t . The *influence graph* at epoch t is denoted by $G_t(V, E, p^{x_t})$, where $p^{x_t} := \{p_{i,j}^{x_t}\}_{(i,j) \in E}$ is the set of influence probabilities and $p_{i,j}^{x_t} \in [0, 1]$ denotes the probability that node i influences node j when the context is x_t . These influence probabilities are unknown to the learner a priori.

At the beginning of epoch t , the learner exogenously influences $k < n$ nodes in the network. The set of these nodes is denoted by S_t , which is also called the *action* at epoch t . An action is an element of the set of k -element subsets of V , which is denoted by $\mathcal{M}$. Nodes in S_t influence other nodes according to the IC model. A node that has not been influenced yet is called an *inactive* node, whereas a node that has been influenced is called an *active* node. In the IC model, each epoch consists of a sequence of time slots indexed by $s \in \{1, 2, \dots\}$. Let $\mathcal{A}_t^s$ denote the set of nodes that are already active at the beginning of time slot s of epoch t , $\mathcal{R}_t^s$ denote the set of nodes that are activated for the first time at time slot s of epoch t , and $\mathcal{C}_t^s$ denote the set of nodes that might be activated at time

slot s of epoch t . In the IC model, we have $\mathcal{A}_t^1 = \emptyset$, $\mathcal{R}_t^1 = S_t$, $\mathcal{A}_t^{s+1} = \mathcal{A}_t^s \cup \mathcal{R}_t^s$ and $\mathcal{C}_t^{s+1} = \{j \in \{\cup_{i \in \mathcal{R}_t^s} \mathcal{N}_i\} - \mathcal{A}_t^{s+1}\}$. For $j \in \mathcal{C}_t^{s+1}$, let $\tilde{\mathcal{V}}_t^{s+1}(j) = \{i \in \mathcal{V}_j \cap \mathcal{R}_t^s\}$ denote the set of nodes in $\mathcal{R}_t^s$ that can influence j . In the IC model, we have

$$\Pr(j \in \mathcal{R}_t^{s+1} | j \in \mathcal{C}_t^{s+1}) = 1 - \prod_{i \in \tilde{\mathcal{V}}_t^{s+1}(j)} (1 - p_{i,j}^{x_t}). \quad (3.2)$$

Suppose that the influence spread process started from a seed set S of nodes. We denote the expected number of endogenously influenced nodes (also called the *expected influence spread*) given context $x \in \mathcal{X}$ and action S as $\sigma(x, S)$, where the expectation is taken over the randomness of the influence given S .

We assume that similar contexts have similar effects on the influence probabilities. This similarity is formalized in the following assumption.

Assumption 3. *There exists $L > 0$, $\theta > 0$ such that for all $(i, j) \in E$ and $x \in \mathcal{X}$, $|p_{i,j}^{x'} - p_{i,j}^x| \leq L\|x' - x\|^\theta$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$.*

3.2.2 Definition of the Reward and the Regret

For a given network graph $G(V, E)$, let $\hat{\mathbf{p}} = \{\hat{p}^x\}_{x \in \mathcal{X}}$ denote the set of estimated and $\mathbf{p} = \{p^x\}_{x \in \mathcal{X}}$ denote the set of true influence probabilities. We define $\hat{\sigma}(x, S)$ as the expected influence spread of action S on $G(V, E, \hat{p}^x)$. For the influence spread process that results from action S , we call an edge $(i, j) \in E$ *activated* if node i influenced node j in this process.

We will compare the performance of the learner with the performance of an oracle that knows the influence probabilities perfectly. For this, we define below the *omnipotent oracle*.

Definition 1. *The omnipotent oracle knows the influence probabilities $p_{i,j}^x$, $\forall (i, j) \in E$ and $\forall x \in \mathcal{X}$. Given context x , it chooses $S^*(x) \in \arg \max_{S \in \mathcal{M}} \sigma(x, S)$ as the*

seed set.

The expected total reward of the omnipotent oracle by epoch T

$$\text{Rew}^*(T) := \sum_{t=1}^T \sigma(x_t, S^*(x_t)).$$

Since finding $S^*(x_t)$ is computationally intractable [15], we propose another (weaker) oracle that only has an approximation guarantee, which is called the (α, β) -approximation oracle ($0 < \alpha, \beta < 1$).

Definition 2. *The (α, β) -approximation oracle knows the influence probabilities $p_{i,j}^x$, $\forall (i, j) \in E$ and $\forall x \in \mathcal{X}$. Given x , it generates an α -approximate solution with probability at least β , i.e., it chooses the seed set $S^{(\alpha, \beta)}(x)$ from the set of actions $\mathcal{M}$ such that $\sigma(x, S^{(\alpha, \beta)}(x)) \geq \alpha \times \sigma(x, S^*(x))$ with probability at least β .*

Note that the expected total reward of the (α, β) -approximation oracle by epoch T is at least $\alpha\beta \times \text{Rew}^*(T)$. Next, we define the approximation algorithm that is used by the learner, which takes the set of estimated influence probabilities as input. Examples of approximation algorithms for the IM problem can be found in [15] and [16].

Definition 3. *The (α, β) -approximation algorithm takes as input the estimated influence probabilities $\hat{p}_{i,j}^x$, $\forall (i, j) \in E$ and $\forall x \in \mathcal{X}$. Given x , it chooses $\hat{S}^{(\alpha, \beta)}(x)$ from the set of actions $\mathcal{M}$ such that $\hat{\sigma}(x, \hat{S}^{(\alpha, \beta)}(x)) \geq \alpha \times \hat{\sigma}(x, \hat{S}^*(x))$ with probability at least β , where $\hat{S}^*(x) \in \arg \max_{S \in \mathcal{M}} \hat{\sigma}(x, S)$.*

Hence, for a sequence of context arrivals $\{x_t\}_{t=1}^T$ the (α, β) -regret of the learner that uses learning algorithm π , which chooses the sequence of actions $\{S_t\}_{t=1}^T$, with respect to the (α, β) -approximation oracle by epoch T is defined as

$$R_\pi^{(\alpha, \beta)}(T) := \alpha\beta \text{Rew}^*(T) - \sum_{t=1}^T \sigma(x_t, S_t) + c \sum_{t=1}^T B_t \quad (3.3)$$

where B_t represents the number of observations in epoch t and c represents the cost per observation.

Our goal is to design online learning algorithms that can work together with any approximation algorithm designed for the offline IM problem, whose expected (α, β) -regrets, i.e., $\mathbb{E}[R_{\pi}^{(\alpha, \beta)}(T)]$, grow slowly in time and in the cardinality of the action set, without making any statistical assumptions on the context arrival process.



Chapter 4

Algorithms

In this section, we study learning algorithms for CMAB with PTAs and OCIMP-CO.

4.1 Algorithms for CMAB with PTAs

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.

4.1.1 CUCB- κ

The pseudocode of CUCB- κ is given in Algorithm 1. CUCB- κ is almost the same as CUCB algorithm given in [24], 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 $\bar{\mu}_i^\kappa$, respectively.

Algorithm 1 Combinatorial UCB- κ (CUCB- κ)

```

1: Input: Set of actions  $\mathcal{S}$ ,  $\kappa > 0$ 
2: Initialize counters: For each arm  $i \in \{1, \dots, m\}$ , set  $T_i = 0$ , which is the
   number of times arm  $i$  is observed.  $t = 1$ 
3: Initialize estimates: Set  $\bar{\mu}_i^\kappa = 1$  and  $\hat{\mu}_i = 1$ ,  $\forall i \in \{1, \dots, m\}$ , which are the
   UCB and sample mean estimates for  $\mu_i$ , respectively
4: while  $t \geq 1$  do
5:   Call the  $(\alpha, \beta)$ -approximation algorithm with  $\bar{\mu}^\kappa$  as input to get  $S_t$ 
6:   Select action  $S_t$ , observe  $X_i^{(t)}$ 's for  $i \in \tau_t$  and collect the reward  $R$ 
7:   for  $i \in \tau_t$  do
8:      $T_i = T_i + 1$ 
9:      $\hat{\mu}_i = \hat{\mu}_i + \frac{X_i^{(t)} - \hat{\mu}_i}{T_i}$ 
10:  end for
11:  for  $i \in \{1, \dots, m\}$  do
12:     $\bar{\mu}_i^\kappa = \min \left\{ \hat{\mu}_i + \kappa \sqrt{\frac{3 \ln t}{2T_i}}, 1 \right\}$ 
13:  end for
14:   $t = t + 1$ 
15: end while

```

Let $\hat{\boldsymbol{\mu}} := \{\hat{\mu}_1, \dots, \hat{\mu}_m\}$ and $\bar{\boldsymbol{\mu}}^\kappa := \{\bar{\mu}_1^\kappa, \dots, \bar{\mu}_m^\kappa\}$ 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 $\bar{\mu}_i^\kappa = \min \left\{ \hat{\mu}_i + \kappa \sqrt{\frac{3 \ln t}{2T_i}}, 1 \right\}$. Initially, CUCB- κ sets T_i^1 as 0 and $\bar{\mu}_i^{\kappa,1}$ as 1 for all arms. Then, in each epoch $t \geq 1$, it calls an (α, β) -approximation algorithm, which takes as input $\bar{\boldsymbol{\mu}}^{\kappa,t}$ and chooses an action S_t . The action S_t depends on the randomness of the approximation algorithm itself in addition to $\bar{\boldsymbol{\mu}}^{\kappa,t}$. 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 , $\mathbf{X}^{(t)}$ and τ_t is collected by the learner. Then, CUCB- κ updates its estimates $\hat{\boldsymbol{\mu}}^{t+1}$ and $\bar{\boldsymbol{\mu}}^{\kappa,t+1}$ 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.

4.1.2 CTS

Algorithm 2 CTS

```

1: Input: Set of actions  $\mathcal{S}$ 
2: Initialize counters: For each arm  $i \in \{1, \dots, m\}$ , set  $s_i = 0$  and  $f_i = 0$ ,
   which are the success and failure counts of arm  $i$ .  $t = 1$ 
3: while  $t \geq 1$  do
4:   For each arm  $i \in \{1, 2, \dots, m\}$ , sample  $\nu_i$  from the  $\text{Beta}(s_i + 1, f_i + 1)$ 
     distribution.
5:   Call the  $(\alpha, \beta)$ -approximation algorithm with  $\boldsymbol{\nu}$  as input to get  $S_t$ .
6:   Select action  $S_t$ , observe  $X_i^{(t)}$ 's for  $i \in \tau_t$  and collect the reward  $R$ .
7:   for  $i \in \tau_t$  do
8:     if  $X_i^{(t)} = 1$  then
9:        $s_i = s_i + 1$ 
10:    else
11:       $f_i = f_i + 1$ 
12:    end if
13:  end for
14:   $t = t + 1$ 
15: 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 3.1, we also assume that $X_i^{(t)} \in \{0, 1\}$ for all $i \in \{1, \dots, m\}$, 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 $i \in \{1, \dots, m\}$, by performing a Bernoulli trial for any arm $i \in \tau_t$ with success probability $X_i^{(t)}$, in a way similar to the extension described in [44]. Also note that Bernoulli arm states are very common, and appear in the OIM problem and the movie recommendation example that we discuss in Section 6.2.

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 $\boldsymbol{\nu} := \{\nu_1, \dots, \nu_m\}$, 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 $i \in \{1, \dots, m\}$, 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 . Note that the algorithm performs this update process as if the arm outcomes ($X_i^{(t)}$'s) are i.i.d for any epoch. However, we do not need this assumption for our analysis to hold. 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.

4.2 Algorithms for Online Contextual Influence Maximization Problem

We propose two algorithms for OCIMP-CO: COIN-CO-EL and COIN-CO-NL. COIN-CO-EL works for the costly edge-level feedback whereas COIN-CO-NL work for the node-level feedback case. Both algorithms use offline IM algorithms as a subroutine, which takes as input the estimated influence probabilities and generate the seed set to be used as action. They, then, update their influence probability estimates in each epoch based on the feedback they gather as a result of their choices of seed set.

4.2.1 Contextual Online Influence Maximization with Costly Edge-level Feedback (COIN-CO-EL)

In this section, we propose the *Contextual Online Influence maximization COstly Edge-Level feedback* (COIN-CO-EL) algorithm. The pseudocode of COIN-CO-EL is given in Algorithm 3. COIN-CO-EL is an online algorithm that can use any offline IM algorithm as a subroutine. In order to exploit the context information efficiently, COIN-CO-EL aggregates information gained from past epochs with similar contexts together in order to form the influence probability estimates. This aggregation is performed by creating a partition $\mathcal{Q}$ of the context set $\mathcal{X}$ based on the similarity information given in Assumption 3. Each set in the partition has a radius that is less than a time-horizon dependent threshold. This implies that the influence probability estimates formed by observations in a certain set of the partition do not deviate too much from the actual influence probabilities that correspond to the contexts that are in the same set.

Recall from (3.2) that in the IC model, at each time slot $s + 1$ of epoch t , nodes in $\mathcal{R}_t^s$ attempt to influence their children by activating the edges connecting them to their children. We call such an attempt in any time slot of epoch t an *activation attempt*. Let F_t be the set of edges with activation attempts at epoch t . F_t is simply the collection of outgoing edges from the active nodes at the end of epoch t , and hence, is known by the learner in the costly edge-level feedback setting. For $(i, j) \in F_t$, we call $a_{i,j}$ the *influence outcome* on edge (i, j) : $a_{i,j} = 1$ implies that node j is influenced by node i while $a_{i,j} = 0$ implies that node j is not influenced by node i . The learner does not have access to $a_{i,j}$'s beforehand, but can observe them by paying a cost c for each observation.¹ COIN-CO-EL keeps two counters $f_{i,j}^Q(t)$ and $s_{i,j}^Q(t)$ for each $(i, j) \in E$ and each $Q \in \mathcal{Q}$. The former denotes the number of observed failed activation attempts on edge (i, j) at epochs prior to epoch t when the context was in Q , while the latter denotes the number of observed successful activation attempts on edge (i, j) in epochs prior to epoch t when the context was in Q .

¹For example, in viral marketing, the marketer can freely observe a person who bought a product [23]. It can also observe the people who influenced that person to buy the product by performing a costly investigation (e.g., conducting a survey).

Algorithm 3 COIN-CO-EL

Require: $T, q_T, G = (V, E), D(t), t = 1, \dots, T$

Initialize sets: Create the partition $\mathcal{Q}$ of $\mathcal{X}$ such that $\mathcal{X}$ is divided into q_T^d identical hypercubes with edge lengths $1/q_T$.

Initialize counters: $f_{i,j}^Q = 0, \forall Q \in \mathcal{Q}, s_{i,j}^Q = 0, \forall Q \in \mathcal{Q}, t = 1$

Initialize estimates: $\hat{p}_{i,j}^Q = 0, \forall (i, j) \in E, \forall Q \in \mathcal{Q}$

```

1: while  $t \leq T$  do
2:   Find the partition  $Q_t \in \mathcal{Q}$  that  $x_t$  belongs to
3:   Compute the set of under-explored edges  $Y_{Q_t}(t)$  given in (4.1) and the set of under-explored
   nodes  $U_{Q_t}(t)$  given in (4.2)
4:   if  $|U_{Q_t}(t)| \geq k$  then {Explore}
5:     Select  $S_t$  randomly from  $U_{Q_t}(t)$ , such that,  $|S_t| = k$ 
6:   else if  $U_{Q_t}(t) \neq \emptyset$  and  $|U_{Q_t}(t)| < k$  then
7:     Select the  $|U_{Q_t}(t)|$  many elements of  $S_t$  as  $U_{Q_t}(t)$  and the remaining  $k - |U_{Q_t}(t)|$  elements of
      $S_t$  by using an  $(\alpha, \beta)$ -approximation algorithm  $G(V, E, \hat{p}_t)$ 
8:   else {Exploit}
9:     Select  $S_t$  by using an  $(\alpha, \beta)$ -approximation algorithm for the IM problem on  $G(V, E, \hat{p}_t)$ 
10:  end if
11:  Observe the set of edges in  $Y_{Q_t}(t) \cap F_t$ , incur cost  $c \times |Y_{Q_t}(t) \cap F_t|$ 
12:  Update the successes and failures  $\forall (i, j) \in Y_{Q_t}(t) \cap F_t$ :
13:  for  $(i, j) \in Y_{Q_t}(t) \cap F_t$  do
14:    if  $a_{i,j} = 1$  then
15:       $s_{i,j}^{Q_t}++$ 
16:    else if  $a_{i,j} = 0$  then
17:       $f_{i,j}^{Q_t}++$ 
18:    end if
19:     $\hat{p}_{i,j}^{Q_t} = \frac{s_{i,j}^{Q_t}}{s_{i,j}^{Q_t} + f_{i,j}^{Q_t}}$ 
20:  end for
21:   $t = t + 1$ 
22: end while

```

At the beginning of epoch t , COIN-CO-EL observes x_t and finds the set $Q \in \mathcal{Q}$ that contains x_t , which is denoted by Q_t .² For each $Q \in \mathcal{Q}$, COIN-CO-EL keeps sample mean estimates of the influence probabilities. For any $x \in Q$ and $(i, j) \in E$, the estimate of $p_{i,j}^x$ at epoch t is denoted by $\hat{p}_{i,j}^Q(t)$.³ This estimate is updated whenever the influence on edge (i, j) is observed by COIN-CO-EL for some context $x \in Q$.

COIN-CO-EL decides on which seed set of nodes S_t to choose based on $\hat{\mathbf{p}}_t := \{\hat{p}_{i,j}^{Q_t}(t)\}_{(i,j) \in E}$. Since these values are noisy estimates of the true influence probabilities, two factors play a role in the accuracy of these estimates: estimation error and approximation error. Estimation error is due to the noise introduced by the randomness of the influence samples, and decreases with the number of samples that are used to estimate the influence probabilities. On the other hand, approximation error is due to the noise introduced by quantization of the context set, and increases with the radius of Q_t . There is an inherent tradeoff between these errors. In order to decrease the approximation error, partition $\mathcal{Q}$ must be refined. This will create more sets in $\mathcal{Q}$, and hence, will result in smaller number of samples in each set, which will cause the estimation error to increase. In order to optimally balance these errors, size of the sets in $\mathcal{Q}$ and the number of observations that fall into each of these sets must be adjusted carefully. COIN-CO-EL achieves this by using a time-horizon dependent partitioning parameter q_T , which is used to partition $\mathcal{X}$ into q_T^d identical hypercubes with border lengths $1/q_T$.⁴ When $\hat{\mathbf{p}}_t$ is far away from $\mathbf{p}$, the estimation accuracy is low. Hence, in order to achieve sublinear regret, the estimate $\hat{\mathbf{p}}_t$ should improve over epochs for all edges $(i, j) \in E$ and for all $Q \in \mathcal{Q}$. This is achieved by alternating between two phases of operation: *exploration* and *exploitation*.

In order to define when COIN-CO-EL explores and exploits, we first define the

²If there are multiple such sets, then one of them is randomly selected.

³We will drop the epoch index when it is clear from the context.

⁴The value of q_T given in Theorem 10 achieves the balance between estimation and approximation errors. When the time horizon T is not known in advance, the same regret bound can be achieved by COIN-CO-EL by using the standard doubling trick [59].

set of under-explored edges at epoch t for $Q \in \mathcal{Q}$, which is given as

$$Y_Q(t) := \{(i, j) \in E \mid f_{i,j}^Q(t) + s_{i,j}^Q(t) < D(t)\} \quad (4.1)$$

where $D(t)$ is a positive, increasing function called the *control function*.⁵ Based on this, the set of under-explored nodes at epoch t for $Q \in \mathcal{Q}$ is defined as

$$U_Q(t) := \{i \in V \mid \exists j \in \mathcal{N}_i : f_{i,j}^Q(t) + s_{i,j}^Q(t) < D(t)\}. \quad (4.2)$$

COIN-CO-EL ensures that the influence probability estimates are accurate when $U_{Q_t}(t) = \emptyset$. In this case, it *exploits* by running an (α, β) -approximation algorithm on $G(V, E, \hat{\mathbf{p}}_t)$. Since, $\hat{\mathbf{p}}_t$ is accurate, it does not pay to observe the influence outcomes on the edges in this phase.

On the other hand, COIN-CO-EL assumes that the influence probability estimates are inaccurate when $U_{Q_t}(t) \neq \emptyset$. In this case, it *explores* by selecting the seed set of nodes according to the following rule: (i) When $|U_{Q_t}(t)| < k$, it selects all of the nodes in $U_{Q_t}(t)$ and the remaining $k - |U_{Q_t}(t)|$ nodes are selected by the (α, β) -approximation algorithm; (ii) When $|U_{Q_t}(t)| \geq k$, k nodes are randomly selected from $U_{Q_t}(t)$. When it explores, COIN-CO-EL also observes the influence outcomes to improve its estimates. For this, it pays and observes the influence outcomes on the edges in the set $Y_{Q_t}(t) \cap F_t$, which denotes the set of under-explored edges with activation attempts.

4.2.2 Contextual Online Influence Maximization with Costly Node-level Feedback (COIN-CO-NL)

The node-level feedback setting is proposed in [23]. In this setting, at the end of an epoch, the learner observes the set of activated nodes, but not the influence outcomes (i.e., edge-level feedback). In this section, we consider an extension to the node-level feedback setting, where at the end of a time slot of an epoch, the learner may choose to observe whether a node is activated or not by paying c for

⁵ $D(t)$ is a sublinear function of t and is also inversely proportional to c .

each observation, which implies that the learner can observe the influence spread process at node-level by costly observations. This is a plausible alternative to the original node-level feedback setting when monitoring status of the nodes in the network is costly. Moreover, obtaining temporal information about when a node gets activated is also plausible in many applications. For instance, in Twitter, a node gets activated when it re-tweets the content of another node that it is following. Similarly, in viral marketing, a node gets activated when it purchases the marketed product. Similar to the previous setting, the goal of the learner is to minimize expectation of the regret given in (3.3). For this purpose, we propose a variant of COIN-CO-EL called COIN-CO-NL, which is able to achieve sublinear regret when only costly node-level feedback is available.

The only difference of COIN-CO-NL from COIN-CO-EL is in the exploration phases. In exploration phases COIN-CO-NL selects the seed set S_t and the nodes to observe $Z_{t,s}$ in time slot s of epoch t in a way that allows perfect inference of influence outcomes on certain edges of the network. We introduce more flexibility to COIN-CO-NL and allow $|S_t| \leq k$. We use the fact that the learner is able to perfectly obtain edge-level feedback from node-level feedback when the children nodes of the seed nodes are distinct. In this case, by observing the children nodes of the seed nodes at $s = 2$ (seed nodes are activated at $s = 1$), the learner can perfectly infer (observe) the influence outcome on the edges between the children nodes and the seed nodes. In order to ensure that the children nodes of the seed nodes are distinct, in the worst-case, the learner can choose a single seed node in exploration phases.

As in COIN-CO-EL, COIN-CO-NL keeps counters $f_{i,j}^Q(t)$ and $s_{i,j}^Q(t)$ for the failed and successful activation attempts perfectly inferred from node-level feedback. These are used at each epoch to calculate $Y_Q(t)$ in (4.1) and $U_Q(t)$ in (4.2). When $U_{Q_t}(t) = \emptyset$, COIN-CO-NL operates in the same way as COIN-CO-EL. When $U_{Q_t}(t) \neq \emptyset$, COIN-CO-NL explores in the following way: When $|U_{Q_t}(t)| \geq k$, instead of choosing k nodes randomly from $|U_{Q_t}(t)|$, it randomly chooses as many nodes as possible from $|U_{Q_t}(t)|$ with distinct children such that $(|\bigcup_{i \in S_t} \mathcal{N}_i| = \sum_{i \in S_t} |\mathcal{N}_i|)$. Similarly, when $|U_{Q_t}(t)| < k$, it randomly chooses as many nodes as possible from $U_{Q_t}(t)$ as long as children nodes of the chosen nodes

are distinct, and chooses the remaining nodes from $V - U_{Q_t}(t)$ as long as $|S_t| \leq k$ and $|\bigcup_{i \in S_t} \mathcal{N}_i| = \sum_{i \in S_t} |\mathcal{N}_i|$. Then, after the seed nodes are chosen it observes all of the nodes j such that $j \in \bigcup_{i \in S_t} \mathcal{N}_i$ and $(i, j) \in Y_{Q_t}(t)$ for some $i \in S_t$ at $s = 2$. This way, $f_{i,j}^Q(t)$ and $s_{i,j}^Q(t)$ values are updated for a subset of the nodes in $Y_{Q_t}(t)$.



Chapter 5

Regret Analysis

In this chapter, we find theoretical performance guarantees for CUCB- κ and CTS, which operate under the framework of CMAB with PTAs as well as for novel algorithms COIN-CO-EL and COIN-CO-NL which operate under OCIMP-CO. The performance guarantees we prove for CUCB- κ and CTS improve the prior work which study algorithms under a more general problem framework. On the other hand, for COIN-CO-EL and COIN-CO-NL, we prove that the regret is sublinear in cost and the number of epochs, and show that our theoretical performance guarantees are the best one can achieve for the non-contextual variant of OCIMP-CO. Theoretical performance guarantees are important because, following the Influence Maximization example, the graph topology as well as the distribution of influence probabilities over the edges can significantly vary, which makes it impractical to perform an exhaustive experimentation.

5.1 Regret Analysis of Algorithms for CMAB with PTAs

In this section we analyze the regrets of CUCB- κ and CTS. Before delving into the details of the regret analysis, we first present 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 large t .

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*

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

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 αr_{μ}^* . 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 l th smallest expected reward. The set of bad actions is denoted by $\mathcal{S}_B := \{S_{B,1}, S_{B,2}, \dots, S_{B,n_B}\}$. Let $\nabla_l := \alpha r_{\mu}^* - r_{\mu}(S_{B,l})$ for each $l \in \{1, \dots, n_B\}$ and $\nabla_{n_B+1} = 0$. Accordingly, we let $\nabla_{\max} := \nabla_1$, $\nabla_{\min} := \nabla_{n_B}$. We also let $\text{gap}(S_t) := \alpha r_{\mu}^* - r_{\mu}(S_t)$.

5.1.1 Regret Analysis for CUCB- κ

First, we show that, given any constant $\delta > 0$, the probability that $\Delta_t^{\kappa} := \max_{i \in \{1, \dots, m\}} |\mu_i - \bar{\mu}_i^{\kappa, t}| < \delta$ 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- κ 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 6.2 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., $\forall T \geq 1$

$$\text{Reg}_{\mu, \alpha, \beta}^{\text{CUCB-}\kappa}(T) \leq \nabla_{\max} \inf_{\eta \in (0, 1)} \left(\lceil t_1 \rceil + \frac{m\pi^2}{3} \left(\frac{2}{\delta^2} + \frac{3}{2} \right) + 2m \left(1 + \frac{2}{\delta^2 \eta p^*} \right) \right) \quad (5.1)$$

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., $\forall T \geq 1$

$$\text{Reg}_{\mu, \alpha, \beta}^{\text{CUCB-0}}(T) \leq \nabla_{\max} \inf_{\eta \in (0, 1)} \left(\lceil t' \rceil + \frac{m\pi^2}{3} \left(1 + \frac{1}{2\delta^2} \right) + 2m \left(1 + \frac{1}{2\delta^2 \eta p^*} \right) \right) \quad (5.2)$$

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 [7, 24, 38] 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 3, 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 smoothness 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 [24] 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 $\text{gap}(S_t) = \alpha r_{\mu}^* - r_{\mu}(S_t)$ based on the change in Δ_t^0 . For this, we use Theorem 2 to bound the expected value of Δ_t^0 , 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*

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

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.

5.1.2 Regret Analysis for CTS

The analysis of the regret of CTS is similar to the regret analysis for CUUCB- κ . We first show that the probability that $\Delta_t^\nu := \max_i |\nu_i^t - \mu_i|$ 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 $t \geq t' := 4c^2/e^2$*

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

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., $\forall T \geq 1$*

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

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 $f(x) =$*

γx^ω where $\gamma > 0$ and $\omega \in (0, 1]$, the gap-independent regret bound for CTS is

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

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 $O(\sqrt{T})$.

5.2 Regret Analysis of Algorithms for Online Contextual Influence Maximization Problem with Costly Observations(OCIMP-CO)

We use the following general framework for the regret analysis for COIN-CO-NL and COIN-CO-EL: First, we find a lower bound for the expected influence spread of (α, β) -approximation algorithms for the Influence Maximization problem. This lower bound is dependent on the accuracy of influence probability estimates; therefore, next, in order to find a regret upper bound for the exploitation epochs, we find an upper bound for the accuracy of the influence probability estimates. This upper bound is achievable because the number of observations is greater than $D(t) = c^{-1/3}t^{2/3}$ by epoch t when t is an exploitation epoch. On the other hand, we find a regret bound for the exploration epochs by upper bounding the number of exploration epochs. This works because we know that number of observations COIN-CO-EL and COIN-CO-NL make is lower bounded by some constant. Hence, coupling the exploitation regret with the exploration regret bound, we achieve a regret upper bound that is sublinear. Finally, we prove a regret lower bound for the non-contextual case that matches the upper regret bound in orders of cost and number of epochs. This suggests that the upper regret bound we find is the best one can achieve for the non-contextual case.

5.2.1 Approximation Guarantee

The maximum difference between the true and estimated influence probabilities given context x is defined as $\Delta_x(\mathbf{p}, \hat{\mathbf{p}}) := \max_{(i,j) \in E} |p_{i,j}^x - \hat{p}_{i,j}^x|$, and the maximum difference over all contexts is defined as $\Delta(\mathbf{p}, \hat{\mathbf{p}}) = \sup_{x \in \mathcal{X}} \Delta_x(\mathbf{p}, \hat{\mathbf{p}})$. The following theorem, originally given as Lemma 6 in [24] provides a relation between the influence spread of action S for $G(V, E, \hat{p}^x)$ and $G(V, E, p^x)$.

Theorem 8. (*Lemma 6 in [24]*) *If $\Delta_x(\mathbf{p}, \hat{\mathbf{p}}) = \Delta_x$, then $|\hat{\sigma}(x, S) - \sigma(x, S)| \leq mn\Delta_x$, for all $S \in \mathcal{M}$.*

The next theorem provides an approximation guarantee for the (α, β) -approximation algorithm with respect to the omnipotent oracle, when it runs using $\hat{\mathbf{p}}$ instead of $\mathbf{p}$.

Theorem 9. *If $\Delta(\mathbf{p}, \hat{\mathbf{p}}) = \Delta$, then*

$$\mathbb{E} \left[\sigma(x, \hat{S}^{(\alpha, \beta)}(x)) \right] \geq \alpha\beta \times \sigma(x, S^*(x)) - \beta(1 + \alpha)mn\Delta$$

for all $x \in \mathcal{X}$.

Proof. See Appendix A.2.1. □

5.2.2 Regret analysis for COIN-CO-EL

The following theorem shows that the expected (α, β) -regret of COIN-CO-EL is sublinear in time for any sequence of context arrivals $x_1, \dots, x_T$. Specifically, when an (α, β) -approximation algorithm is used as the offline IM algorithm in COIN-CO-EL, then the expectation of the regret of COIN-CO-EL given in (3.3) is bounded by a sublinear function of the number of epochs. This implies that the expected regret of COIN-CO-EL averaged over epochs converges to zero as the number of epochs increases, and hence, the performance of COIN-CO-EL converges to the performance of the omnipotent oracle.

Theorem 10. *When COIN-CO-EL uses an (α, β) -approximation algorithm as the subroutine, and when $q_T = \lceil T^{1/(3\theta+d)} \rceil$ and $D(t) = (c+1)^{-2/3} t^{2\theta/(3\theta+d)}$, we have*

$$\begin{aligned}
& \mathbb{E}[R_{\text{COIN-CO-EL}}^{(\alpha, \beta)}(T)] \\
& \leq \left(\frac{m}{k} + 1\right) \alpha \beta (n - k) \left\lceil T^{\frac{1}{3\theta+d}} \right\rceil^d \left\lceil (c+1)^{-2/3} T^{\frac{2\theta}{3\theta+d}} \right\rceil \\
& \quad + mc \left\lceil (c+1)^{-2/3} T^{\frac{2\theta}{3\theta+d}} \right\rceil \left\lceil T^{\frac{1}{3\theta+d}} \right\rceil^d \\
& \quad + \beta(1 + \alpha) mn L d^{\theta/2} T^{\frac{2\theta+d}{3\theta+d}} \\
& \quad + \frac{\beta(1 + \alpha) \pi m^2 n (c+1)^{1/3}}{\sqrt{2}} \times \frac{T^{\frac{2\theta+d}{3\theta+d}}}{\frac{2\theta+d}{3\theta+d}} \\
& = O\left(c^{1/3} T^{\frac{2\theta+d}{3\theta+d}}\right)
\end{aligned}$$

for an arbitrary sequence of contexts $\{x_t\}_{t=1}^T$.

Proof. See Appendix A.2.2. □

Remark 1. *As observed from Theorem 10, COIN-CO-EL explores less when the cost of observation is large. In addition, when the cost is 0, the regret bound is equivalent to the regret bound in [26], which does not consider the observation costs and assumes that the influence outcomes are always observed. This result shows that sublinear number of observations is sufficient to achieve the same order of regret as in [26].*

The next corollary gives an upper bound on the expected (α, β) -regret of COIN-CO-EL when $\theta > 1$, which corresponds to the context-free online IM problem with costly observations.

Corollary 1. *When $\theta > 1$ and COIN-CO-EL uses an (α, β) -approximation algorithm as the subroutine with $q_T = 1$ and $D(t) = (c+1)^{-2/3} t^{2/3}$, we have*

$$\mathbb{E}[R_{\text{COIN-CO-EL}}^{(\alpha, \beta)}(T)] = O\left(c^{1/3} T^{2/3}\right).$$

Proof. The proof follows directly from the proof of Theorem 10. Since $q_T = 1$,

sum of the regrets incurred over exploration epochs and due to observing the influence outcomes is proportional to $c \times D(T) = O(c^{1/3}T^{2/3})$. Moreover, the regret incurred over an exploitation epoch only depends on the estimation error since there is no approximation error. Essentially, from Lemma 13, it is observed that if the learner exploits at epoch t , then it incurs at most $O((c+1)^{1/3}t^{-1/3})$ regret in that epoch. Summing this from 1 to T gives $O(c^{1/3}T^{2/3})$ regret due to exploitation epochs. $\square$

The regret bounds given in Theorem 10 and Corollary 1 are gap-independent. For the cost-free online IM problem it is shown in [24] that there exists a learning algorithm with $\tilde{O}(T^{1/2})$ gap-independent regret

5.2.3 Regret analysis for COIN-CO-NL

The (α, β) -regret of COIN-CO-NL is bounded in the following theorem.

Theorem 11. *When COIN-CO-NL uses an (α, β) -approximation algorithm as the subroutine, and when $q_T = \lceil T^{1/(3\theta+d)} \rceil$ and $D(t) = (c+1)^{-2/3}t^{2\theta/(3\theta+d)}$, we have*

$$\begin{aligned}
& \mathbb{E}[R_{\text{COIN-CO-NL}}^{(\alpha, \beta)}(T)] \\
& \leq \alpha\beta m(n-k) \left\lceil T^{\frac{1}{3\theta+d}} \right\rceil^d \left\lceil (c+1)^{-2/3} T^{\frac{2\theta}{3\theta+d}} \right\rceil \\
& \quad + mc \left\lceil (c+1)^{-2/3} T^{\frac{2\theta}{3\theta+d}} \right\rceil \left\lceil T^{\frac{1}{3\theta+d}} \right\rceil^d \\
& \quad + \beta(1+\alpha)mnLd^{\theta/2}T^{\frac{2\theta+d}{3\theta+d}} \\
& \quad + \frac{\beta(1+\alpha)\pi m^2 n(c+1)^{1/3}}{\sqrt{2}} \times \frac{T^{\frac{2\theta+d}{3\theta+d}}}{\frac{2\theta+d}{3\theta+d}} \\
& = O\left(c^{1/3}T^{\frac{2\theta+d}{3\theta+d}}\right)
\end{aligned}$$

for an arbitrary sequence of contexts $\{x_t\}_{t=1}^T$.

Proof. See Appendix A.2.2. $\square$

Like COIN-CO-EL, COIN-CO-NL also ensures that when it is in an exploitation epoch t , each edge is observed at least $D(t)$ many times. However, because exploration phases last longer under node-level feedback, the regret incurred due to explorations is greater. In the worst-case, COIN-CO-NL can only update one edge during one exploration epoch.

The next corollary gives an upper bound on the expected (α, β) -regret of COIN-CO-NL when $\theta > 1$, which corresponds to the context-free online IM problem with costly node-level feedback.

Corollary 2. *When $\theta > 1$ and COIN-CO-NL uses an (α, β) -approximation algorithm as the subroutine with $q_T = 1$ and $D(t) = (c + 1)^{-2/3}t^{2/3}$, we have*

$$\mathbb{E}[R_{\text{COIN-CO-NL}}^{(\alpha, \beta)}(T)] = O(c^{1/3}T^{2/3}).$$

Proof. The proof follows directly from the proof of Theorem 11 and the proof of Corollary 1. $\square$

In the following subsection, we show that learning in the online IM problem with costly observations is inherently more difficult than the cost-free version of the same problem by proving $\Omega(c^{1/3}T^{2/3})$ lower bound on the regret. Finally, we note that Theorem 12 in the next section provides a matching lower bound for the node-level feedback setting as well as for the edge-level feedback setting.

5.2.4 Lower Bound on the Regret for the Context-free Online IM Problem

In this section, we consider the special case of OCIMP-CO when $\theta > 1$, and show that the regret lower bound is $\Omega(c^{1/3}T^{2/3})$.

Theorem 12. *Consider OCIMP-CO with $\theta > 1$. Assume that both edge-level and node-level feedbacks are costly, where the cost of each observation is $c > 0$. For this problem, there exists a problem instance (influence graph) for which any*

learning algorithm π run with an exact solver, i.e., $(\alpha, \beta) = (1, 1)$, which makes OT^1 observations by epoch T , will incur regret

$$\begin{aligned} & \mathbb{E}[R_{\pi}^{(1,1)}(T)] \\ & \geq \max \left\{ 1.88 \times \left(\frac{\sqrt{6}}{16} \right)^{2/3} k_0^{2/3} c^{1/3} m^{1/3} T^{2/3}, \frac{\sqrt{6}}{16} k_0 \sqrt{T} \right\} \end{aligned}$$

for $T \geq \frac{3}{8} \frac{m}{O} k_0$, where $k_0 = \left(1 - \frac{k}{m}\right)^2$.

Proof. See Appendix A.2.3. □

Proof of Theorem 12 is built on the lower bound proofs in [35, 60–62] developed for prediction with expert advice and MAB problems. The differences lie in the formulation of the problem instance we show our worst-case regret lower bound on, and the way we handle actions, which do not correspond to individual arms but a combination of the arms. In particular, we use the fact that we can decouple actions from the arms as long as observations of the arms determine the actions taken by the learner.

¹Here, O is the average number of observations made by the algorithm in an epoch, which is a positive real number such that OT is an integer. Since we have at most m observations in each epoch, $0 < O \leq m$.

Chapter 6

Experimental Results

6.1 Online Contextual Influence Maximization Problem with Costly Observations(OCIMP-CO)

First, we explain extensions of algorithms used in experiments. First two extensions are for improving the efficiency of the our proposed algorithms COIN-CO-EL and COIN-CO-NL without sacrificing the theoretical guarantees whereas the third extension is an adaptation of the famous ϵ_t -Greedy algorithm to OCIMP-CO.

6.1.1 Extensions

6.1.1.1 Improved Exploration Phase

To improve the performance of COIN-CO-EL in exploration phases, we consider two additional exploration strategies. In the first variant, which is called COIN-CO-EL+, instead of choosing nodes in S_t randomly from $U_{Q_t}(t)$, a modified version of TIM+ [16], which is restricted to choose S_t only from $U_{Q_t}(t)$, is used to select

S_t . The intuition behind this choice is that since TIM+ is an (α, β) -approximation algorithm, it may provide a larger influence spread, even when the influence probability estimates are not completely accurate. In the second variant, which is called COIN-CO-EL-HD, S_t is chosen using the High-Degree heuristic [21]. High-Degree chooses nodes who have the highest out-degree values as its seed set. The motivation for using High-Degree in the exploration phases is twofold. Firstly, since the influence probability estimates are highly inaccurate in the initial epochs, an (α, β) -approximation algorithm whose performance depends on the accuracy of the influence probability estimates may not work well. Therefore, High-Degree, which does not use these estimates but uses the graph structure can work better. Secondly, High-Degree is much faster than an (α, β) -approximation algorithm, since its node selection strategy is very simple. Moreover, both COIN-CO-EL+ and COIN-CO-EL-HD have the same theoretical performance guarantees as COIN-CO-EL, since the regret analysis carried out for COIN-CO-EL is agnostic to the type of algorithm used to select from the set of under-explored nodes. We perform experiments on COIN-CO-EL+ and COIN-CO-EL-HD in Section 6.1.2.

6.1.1.2 A Randomized Algorithm for Costly Edge-level Feedback: ϵ_t -Greedy-CO-EL

In this section, we propose an algorithm that is inspired by the ϵ_t -greedy strategy for the MAB problem proposed in [35]. This algorithm, called ϵ_t -Greedy-CO-EL (pseudocode given in Algorithm 2), is similar to the online IM algorithm presented in [31]. While it does not take the context into account, in our experiments we run a context-aware version of this algorithm by using the procedure described in Algorithm 5.

As its name suggests, in each epoch ϵ_t -Greedy-CO-EL explores with probability ϵ_t and exploits with probability $1 - \epsilon_t$, where ϵ_t is a positive decreasing function of t . For an edge (i, j) , ϵ_t -Greedy-CO-EL keeps the parameters $s_{i,j}$ and $f_{i,j}$, which are the counters for the observed success and failures, respectively. When it exploits, it uses an (α, β) -approximation algorithm with the sample mean estimates of the influence probabilities as input to select the set of seed nodes. On the other hand,

Algorithm 4 ϵ_t -Greedy-CO-EL

Require: $T, G = (V, E), \epsilon_t, t = 1, \dots, T$

Initialize counters: $f_{i,j} = 0, s_{i,j} = 0, \forall (i, j) \in E, t = 1$

```
1: while  $t \leq T$  do
2:   Sample  $z \sim \text{Bernoulli}(\epsilon_t)$ 
3:   if  $z = 1$  then {Explore}
4:      $\hat{p}_{i,j} = \min \left\{ \frac{s_{i,j}}{s_{i,j} + f_{i,j}} + \sqrt{\frac{s_{i,j} f_{i,j}}{(s_{i,j} + f_{i,j})^2 (s_{i,j} + f_{i,j} + 1)}}, 1 \right\}$ 
5:     Select  $S_t$  by using an  $(\alpha, \beta)$ -approximation algorithm for the IM problem on
        $G(V, E, \{\hat{p}_{i,j}\}_{(i,j) \in E})$ 
6:     Observe the set of edges in  $F_t$ , incur cost  $c \times |F_t|$ 
7:     Update the successes and failures  $\forall (i, j) \in F_t$ :
8:     for  $(i, j) \in F_t$  do
9:       if  $a_{i,j} = 1$  then
10:         $s_{i,j}++$ 
11:       else if  $a_{i,j} = 0$  then
12:         $f_{i,j}++$ 
13:       end if
14:     end for
15:   else {Exploit}
16:      $\hat{p}_{i,j} = \frac{s_{i,j}}{s_{i,j} + f_{i,j}}$ 
17:     Select  $S_t$  by using an  $(\alpha, \beta)$ -approximation algorithm for the IM problem on
        $G(V, E, \{\hat{p}_{i,j}\}_{(i,j) \in E})$ 
18:   end if
19:    $t = t + 1$ 
20: end while
```

when it explores it uses an inflated version of the influence probability estimates, which is given as the sample mean plus the sample standard deviation. Then, it provides these as input to the (α, β) -approximation algorithm to select the seed set of nodes. Using the additional standard deviation term allows it to explore influence outcomes of the edges that are sampled relatively less, working similar to the inflation factor used in UCB algorithms. When ϵ_t -Greedy-CO-EL explores, it observes the influence outcomes on all edges with activation attempts.

6.1.2 Experiments

In this section, we carry out numerical experiments to compare the performance of our algorithms with existing ones in numerous different settings. Besides the algorithms that we introduce in this paper, which work under the costly observation setting, we also run simulations with COIN+ [26] and COIN-HD,

which work under the assumption that observing influence outcomes on the edges and the set of activated nodes have no cost. Hence, these algorithms observe at each epoch the influence outcomes on all edges with activation attempts.

6.1.2.1 Feedback Mechanisms

We consider three different feedback mechanisms in our experiments:

Cost-Free Edge-level Feedback: In this setting, the influence outcomes on all edges with activation attempts are observed at the end of each epoch with no cost ($c = 0$). This is the setting that is used in our prior work [26].

Cost-Free Node-level Feedback: In this setting, all activated nodes are observed at the end of each epoch without paying any observation cost [23]. Besides, there is no other feedback available.

Costly Edge-level Feedback: This is the feedback setting proposed in this paper, which is explained in detail in Section 3.2.

6.1.2.2 Setup

We conducted our experiments in four configurations that is composed of two settings and two social networks. The settings are called contextual and non-contextual. The latter is a special case of OCIMP, in which the influence probabilities are context-independent.

We use a real-world and a synthetic network: NetHEPT and NetHEPT-. NetHEPT is extensively used in IM literature [16, 23, 31], and NetHEPT- is a random subgraph of NetHEPT where all of the nodes have a positive in-degree. In NetHEPT, roughly a third of the nodes have an in-degree value of 0, which means that they cannot be activated endogenously whereas in NetHEPT-, all of the nodes can be activated by a choice of seed set. In our experiments, we set $T = 5000$.

Table 6.1: Properties of the networks used in the experiments.

Dataset	$ V $	$ E $	Average In-degree
NetHEPT	15K	59K	3.86
NetHEPT-	4K	10.5K	2.63

We consider one dimensional contexts, i.e., $d = 1$ and assume that $k = 50$, which is a typical choice in the online IM literature [23,31]. For our algorithms, we set $q_T = 2$ and initialize all influence probability estimates as 0. In order to make exploration phases of the algorithms scalable, for the edge-level and node-level feedback experiments, we set $D(t) = \frac{t^{2/5}}{100}$, and for the costly edge-level feedback, we set $D(t) = \frac{e^{-2/3}t^{2/5}}{200}$ and $c = 0.1$.

We report both the time averaged regret and ℓ_2 -error of the influence probability estimates, where ℓ_2 -error at epoch t is given by $\ell_2^t := \sqrt{\sum_{(i,j) \in E} (p_{i,j}^{x_t} - \hat{p}_{i,j}^{Q_t})^2}$. The (α, β) -approximation algorithm used by our learning algorithms is chosen as TIM+ [16].

6.1.2.3 Defining the Influence Probabilities

The context x_t in any epoch t is sampled uniformly at random from $[0, 1]$. The influence probabilities are generated according to a Hölder-continuous surface over $[0, 1]$ defined by the following equations:

$$\frac{0.89}{1 + e^{(-1000 \times (x_t - 0.5))}} + 0.01, \quad (6.1)$$

$$\frac{0.89}{1 + e^{(-1000 \times ((1 - x_t) - 0.5))}} + 0.01. \quad (6.2)$$

In our simulations, we consider a network composed of two groups of nodes

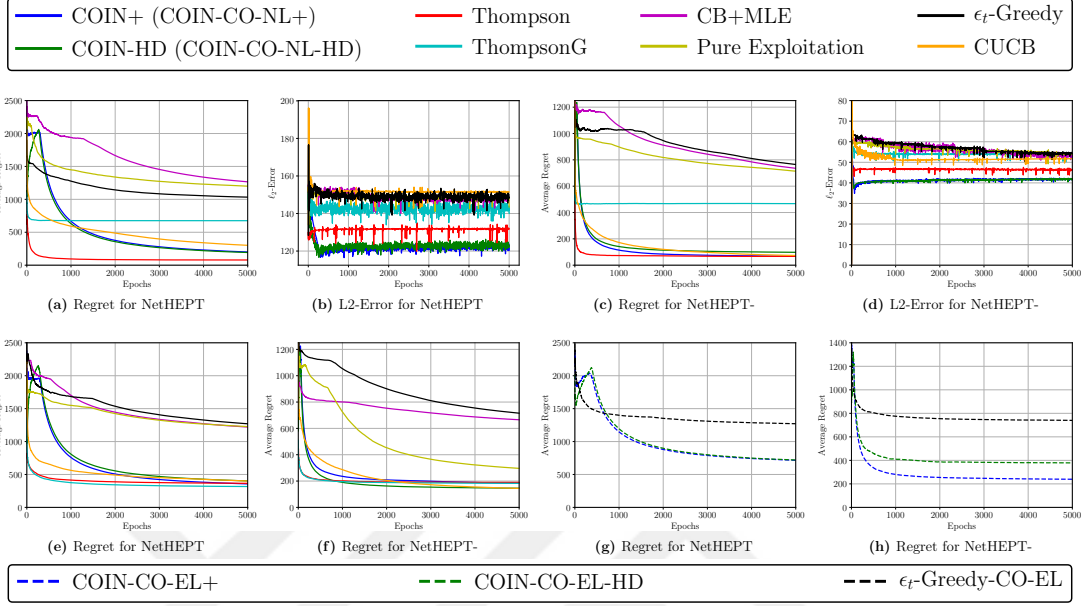


Figure 6.1: Results for context-aware algorithms under cost-free edge-level feedback (a-d), cost-free node-level feedback (e-f) and costly edge-level feedback (g-h).

with conflicting opinions or interests. For this, we randomly partition the nodes in the network into two groups. The influence probabilities of the outgoing edges of the nodes in the two groups are calculated using (6.1) and (6.2) such that the influence probabilities are roughly between 0.01 and 0.9. Hence, when edges in one group has high influence probabilities, the edges in the other group has low influence probabilities.

6.1.2.4 Algorithms for the Cost-free Edge-Level Feedback Setting

We compare the performance of the proposed algorithms with various algorithms that we adapt to the OCIMP framework. Inspired by the structure of COIN-CO-EL, we use the approach presented in Algorithm 3 to adapt a context-free MAB algorithm π to the OCIMP framework. Formally, we expect q_T^d independent instances of the algorithm denoted by $\{\pi_i\}_{i=1}^{q_T^d}$ to be run on separate sets in the context partition. In order to improve the performance in node-level feedback simulations, we also use the extension proposed in Section 6.1.1.1 for node-level feedback.

Algorithm 5 Adapting MAB Algorithms to OCIMP

Require: $\{\pi_i\}_{i=1}^{q_T^d}, T, q_T, G = (V, E)$

- 1: Create the partition $\mathcal{Q}$ of $\mathcal{X}$ such that $\mathcal{X}$ is divided into q_T^d identical hypercubes with edge lengths $1/q_T$
 - 2: Initialize i^{th} instance of the algorithm $\pi_i, \forall i \in \{1, \dots, q_T^d\}$
 - 3: $t = 1$.
 - 4: **while** $t \leq T$ **do**
 - 5: Find the partition $Q_t \in \mathcal{Q}$ that x_t belongs to
 - 6: Get arm indices $\hat{\mathbf{p}}_t = \{p_{i,j}^{Q_t}\}_{(i,j) \in E}$ from π_{Q_t}
 - 7: Select S_t by using an (α, β) -approximation algorithm for the IM problem on $G(V, E, \hat{\mathbf{p}}_t)$
 - 8: Obtain observations using the appropriate feedback setting
 - 9: Update π_{Q_t} using the observations
 - 10: $t = t + 1$
 - 11: **end while**
-

COIN+: A variant of COIN (a contextual learning algorithm for OCIMP proposed in [26]), which utilizes TIM+ to choose the seed set of nodes in exploration phases.

COIN-HD: A variant of COIN, which utilizes the High-Degree heuristic to choose the seed set of nodes in exploration phases.

Thompson: Draws the estimated influence probability of each edge from a Beta distribution, where the parameters of the Beta distribution for edge (i, j) in epoch t are $s_{i,j}$ and $f_{i,j}$, which are the counters for the observed successes and failures for edge (i, j) , and both are initialized as 1. Thompson updates these parameters in each epoch according to edge-level feedback.

ThompsonG: A variant of Thompson, where the parameters of the Beta distribution from which the estimated influence probabilities are drawn are calculated using global priors 1 and 19 for α and β parameters, respectively, as explained in [31]. These global priors are updated in each epoch t according to the edge-level feedback as in [31].

CB+MLE: UCB-based algorithm proposed for the online IM problem that is explained in [31].

ϵ_t -Greedy : ϵ_t -Greedy is a variant of ϵ_t -Greedy-CO-EL, which works for the cost-free case ($c = 0$).

Pure Exploitation: A variant of COIN that does not explore. It only performs implicit exploration due to the probabilistically triggered edges, and works under the edge-level and node-level feedback settings.

CUCB: A UCB-based algorithm proposed in [24] for the combinatorial MAB problem.

6.1.2.5 Algorithms for the Cost-free Node-level Feedback Setting

For this setting, it is known that for each endogenously influenced node $i \in V$, at least one of its parent nodes $j \in V_i$. Thus, we use the frequentist credit assignment method proposed in [23] to adapt an algorithm designed for cost-free edge-level feedback to work under cost-free node-level feedback.

For this, let $\mathcal{V}'_i \subset \mathcal{V}_i$ denote the set of active parents of i . In this method, we assume that the probability with which a node $i \in V$ is influenced by a node $j \in \mathcal{V}'_i$ is $1/|\mathcal{V}'_i|$. Then, we sample from this distribution one of the nodes $l \in \mathcal{V}'_i$ as the influencer of node i , set $a_{l,i} = 1$, and $a_{j,i} = 0$ for all other $j \in \mathcal{V}'_i$. Finally the assigned influence outcomes are used to update the influence probability estimates. All algorithms proposed for the cost-free edge-level feedback setting are adapted using the above procedure to work under the cost-free node-level feedback setting.

6.1.2.6 Algorithms for the Costly Edge-level Feedback Setting

For this setting, we use COIN-CO-EL+, COIN-CO-EL-HD and ϵ_t -Greedy-CO-EL described in Sections 6.1.1.2 and 6.1.1.1. As its learning parameter, ϵ_t -Greedy-CO-EL uses $t = 1/\sqrt{t}$. Since other algorithms do not explicitly separate exploration and exploitation phases, their extension to this setting is not straightforward, and hence, they are not considered for this setting.

6.1.3 Results for Cost-Free Edge-level Feedback

This setting allows cost-free observations and hence corresponds to the special case of $c = 0$ in our framework. We consider the more challenging costly observation setting in Section VIII-G. For cost-free edge-level feedback setting, we compare the performances of COIN-HD and COIN+ [26] with the other algorithms, as shown in Fig. 2.

Regret Comparison. Our experiments demonstrate that most of the other considered algorithms are outperformed by COIN+ and COIN-HD, where only CUCB and Thompson are competitive. Especially in NetHEPT, which is a bigger graph than NetHEPT-, COIN+ and COIN-HD do long explorations in the first epochs where they suffer from high exploration regret. That is why, initially, their average regret is much higher than that of other algorithms. However, in this phase, they discover the influence probabilities of the graph well enough to achieve lower regret in exploitation epochs, and are able to quickly reduce their average regret.

Learning Influence Probabilities. We investigate the ℓ_2 -errors of the considered algorithms to understand how accurate their estimates of influence probabilities are. For both networks, COIN-HD and COIN+ achieve lower ℓ_2 -errors than all other algorithms due to their explicit exploration phases. However, as seen from the average regret results, having accurate influence probability estimates does not always correspond to achieving small average regret. This is due to the fact that knowledge about the influence probabilities associated with nodes that are not influential in the network does not have much impact on the seed set selection process. However, we argue that in more challenging networks of larger size, the disparity between ℓ_2 -errors of these algorithms would be more indicative of their performance as the algorithms that do not conduct extensive exploration phases would be more likely to miss out on some of the more influential nodes in the network, and hence, under-perform.

6.1.4 Results for Cost-Free Node-level Feedback

This setting allows the algorithms to observe node-level feedback without cost and corresponds to the case of $c = 0$ in our formulation. For cost-free node level feedback setting, we present results for COIN-CO-NL+ and COIN-CO-NL-HD along with the listed benchmark algorithms adapted to node level feedback setting. We have used the procedure in Algorithm 3 together with the node-level update described in Section 6.1.2.6. Since the ℓ_2 -error trend in this and the following experiments are similar to the results presented for edge-level feedback, we consider only the regret hereafter for the sake of brevity.

Regret Comparison. For the node-level feedback setting, we observe in Fig. 2 (e-f) that COIN-CO-NL+ and COIN-CO-NL-HD outperform most of the benchmark algorithms and perform on par with CUCB and Thompson in further epochs. Similar to the case for the cost-free edge-level feedback, initially both COIN-CO-NL+ and COIN-CO-NL-HD suffer high average regrets due to their extensive explorations, but benefit from it in long term by achieving low exploitation regret and catching up with the best performing algorithms.

One interesting observation that highlights the importance of extensive explorations is the performance of Pure Exploitation in this setting. The performance of this greedy algorithm is very poor for NetHEPT, but significantly improves for NetHEPT-. The smaller scale of the latter network benefits the algorithm as its implicit exploration while exploiting is enough to discover most of the network. However, in the case of NetHEPT, due to the larger scale of the network and the existence of nodes with zero in-degree, the greediness of the algorithm hurts it as it can not estimate many important influence probabilities accurately.

This observation supports the discussion about the ℓ_2 -error results. The bigger and more complex the network is, the more valuable thorough exploration seems to become for the learning algorithms. We would expect to see the impact of the costly exploration phases of COIN-CO-NL+ and COIN-CO-NL-HD in more challenging settings, where the implicit exploration of CUCB and Thompson might not be able to identify all of the influential nodes.

6.1.5 Results for the Costly Edge-level Feedback

For costly edge-level feedback setting, we compare COIN-CO-EL-HD, and COIN-CO-EL+ with ϵ_t -Greedy-CO-EL. Since the other algorithms do not explicitly separate exploration and exploitation phases, their extension to this setting is not straightforward, and hence, they are not considered in this experiment.

Regret Comparison. We observe in Fig. 2 (g-h) that COIN-CO-EL-HD and COIN-CO-EL+ perform significantly better than ϵ_t -greedy-CO-EL on both networks. We also highlight the difference of regrets incurred over exploration and exploitation phases of COIN-CO-EL+ and COIN-CO-EL-HD in Fig. 6.2. These are calculated separately by taking average of the regrets while the algorithm is in exploration epochs and while it is in exploitation epochs.

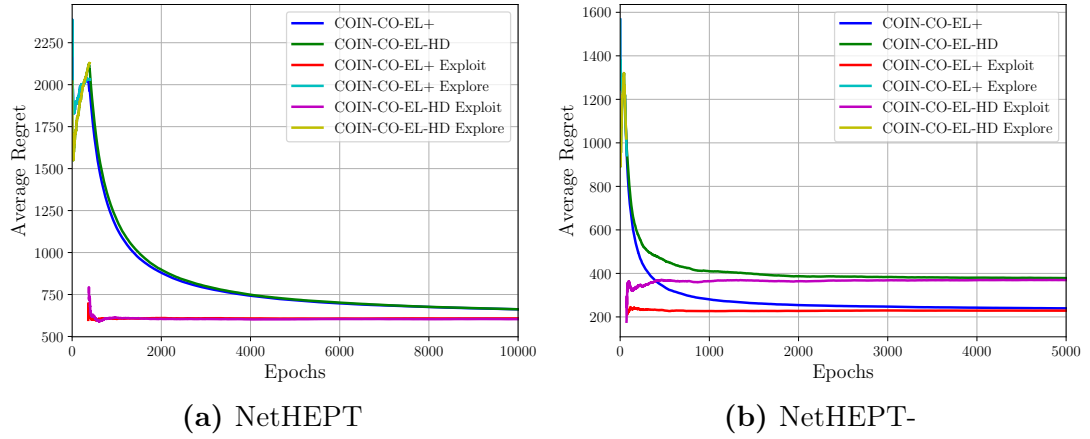


Figure 6.2: Average regrets of COIN-CO-EL+ and COIN-CO-EL-HD calculated over exploration and exploitation epochs.

We observe that immediately after their high regret exploration epochs, COIN-CO-EL+ and COIN-CO-EL-HD start to perform very close to the (α, β) -approximation oracle, suffering substantially lower regrets. The impact of this decrease is reflected on the average regret in the long run.

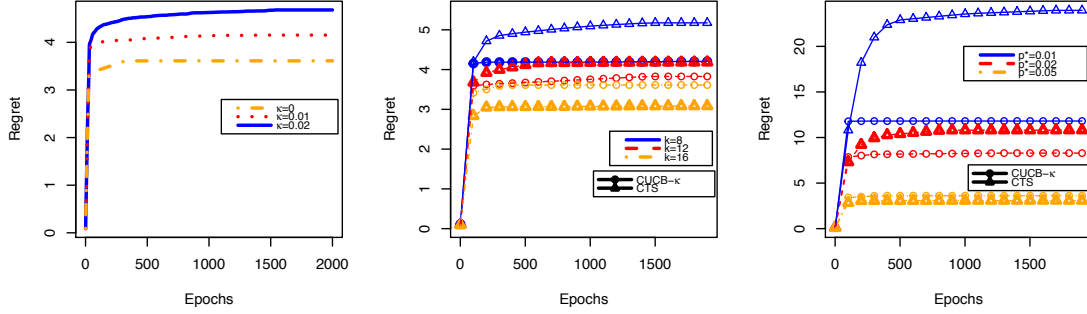


Figure 6.3: Regrets of CUCB- κ and CTS for different parameter values. Left figure: CUCB- κ for $\kappa = 0, 0.01, 0.02$. Middle figure: CUCB- κ for $\kappa = 0$ (circle marker) and CTS (triangle marker). Right figure: CUCB- κ for $\kappa = 0$ (circle marker) and CTS (triangle marker).

6.2 CMAB with PTAs

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.

6.2.1 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

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

maximum coverage problem [24]. 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 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 [63], 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.

6.2.2 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 $g_{i_k} = 1$ 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 $p_{i,j} = sc \times \frac{\langle \mathbf{m}_i, \mathbf{u}_j \rangle r_i}{\max_i r_i}$, $(i, j) \in E$, 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.

6.2.3 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 [63], by giving the true influence probabilities as input. We observe from Figure 6.3 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 also observed that both CUCB- κ and CTS incur 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 .



Chapter 7

Conclusion

In this thesis, we propose a new online influence maximization problem, called OCIMP-CO, where the novelties are that the influence probabilities depend on the context and making observations are costly. We develop computationally efficient learning algorithms for this problem. We prove that they achieve sublinear regret, which implies that the average regret will approach zero and show that our regret bound is the best possible one can achieve in cost and time orders for the non-contextual case of the problem. We also demonstrate that these algorithms perform on par with or better than their competitors on real-world networks.

We also consider a problem that is closely linked with OCIMP-CO, called CMAB with PTAs, where the context and cost of observation are not considered. This problem contains the Online Influence Maximization as a special case. The version of the problem we study has positive Arm Triggering Probabilities(ATPs). For this problem, we 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. These results are improvements over prior work on CMAB with PTAs which does not assume that ATPs are strictly positive [54]. We also numerically show that CUCB- κ and CTS achieve bounded regret in a 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. This is due to the fact that positive ATPs result in implicit exploration which is formalized in Theorem 1. Using the example of the OIM problem, a conclusion of Theorem 1 is the following: Positive ATPs imply that any node will get influenced sufficiently many times with high probability, independently of the algorithm, therefore, independently of the action set used, which may eliminate the need for exploration strategies.

Our regret bound for the CMAB with PTAs includes the $1/p^*$ term that is shown to be unavoidable in [54]. In OIM terms, p^* is the minimum positive probability with which a node is influenced. This term is potentially problematic as $p^* > 0$ can be arbitrarily small, which causes the regret bound to be arbitrarily large. Future work might include regret bounds under a more strict assumption on the expected reward function that avoids the p^* term. An example of such assumption is the triggering probability modulated bounded smoothness condition given in [54]. This assumption is motivated by the OIM problem, and includes the observation that the nodes that are less likely to be triggered have a smaller effect on the expected influence spread in the Influence Maximization problem.

Bibliography

- [1] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays — part i: i.i.d. rewards,” in *IEEE Transactions on Automatic Control*, pages=968-976, year=1987,.
- [2] Y. Gai, B. Krishnamachari, and R. Jain, “Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations,” *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1466–1478, 2012.
- [3] W. Chen, Y. Wang, and Y. Yuan, “Combinatorial multi-armed bandit: General framework and applications,” in *Proc. ICML*, pp. 151–159, 2013.
- [4] A. Gopalan, S. Mannor, and Y. Mansour, “Thompson sampling for complex online problems,” in *Proc. ICML*, vol. 32, pp. 100–108, 2014.
- [5] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson, “Matroid bandits: Fast combinatorial optimization with learning,” *arXiv preprint arXiv:1403.5045*, 2014.
- [6] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan, “Cascading bandits: Learning to rank in the cascade model,” in *Proc. ICML*, pp. 767–776, 2015.
- [7] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, “Combinatorial cascading bandits,” in *Proc. NIPS*, pp. 1450–1458, 2015.

- [8] X. Cao, Y. Chen, C. Jiang, and K. J. R. Liu, “Evolutionary information diffusion over heterogeneous social networks,” *IEEE Trans. Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 595–610, 2016.
- [9] X. Cao, Y. Chen, and K. J. R. Liu, “Understanding popularity dynamics: Decision-making with long-term incentives,” *IEEE Trans. Signal and Information Processing over Networks*, vol. 3, no. 1, pp. 91–103, 2017.
- [10] S. Chen, J. Fan, G. Li, J. Feng, K.-l. Tan, and J. Tang, “Online topic-aware influence maximization,” vol. 8, pp. 666–677, 2015.
- [11] N. Barbieri, F. Bonchi, and G. Manco, “Topic-aware social influence propagation models,” in *Proc. IEEE 12th International Conference on Data Mining (ICDM)*, pp. 81–90, 2012.
- [12] Ç. Aslay, N. Barbieri, F. Bonchi, and R. A. Baeza-Yates, “Online topic-aware influence maximization queries,” in *EDBT*, pp. 295–306, 2014.
- [13] W. Chen, T. Lin, and C. Yang, “Real-time topic-aware influence maximization using preprocessing,” *Computational Social Networks*, vol. 3, no. 1, pp. 1–13, 2016.
- [14] A. O. Saritac, A. Karakurt, and C. Tekin, “Online contextual influence maximization with costly observations,” *IEEE Transactions on Signal and Information Processing over Networks*, 2018. © 2018 IEEE. Reprinted, with permission.
- [15] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in *Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 137–146, 2003.
- [16] Y. Tang, X. Xiao, and Y. Shi, “Influence maximization: Near-optimal time complexity meets practical efficiency,” in *Proc. ACM SIGMOD International Conference on Management of Data*, 2014.
- [17] K. Jung, W. Heo, and W. Chen, “Irie: Scalable and robust influence maximization in social networks,” in *Proc. IEEE 12th International Conference on Data Mining (ICDM)*, pp. 918–923, 2012.

- [18] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,” in *Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 420–429, 2007.
- [19] A. Goyal, W. Lu, and L. V. Lakshmanan, “Celf++: Optimizing the greedy algorithm for influence maximization in social networks,” in *Proc. 20th International Conference Companion on World Wide Web*, pp. 47–48, 2011.
- [20] Y. Tang, Y. Shi, and X. Xiao, “Influence maximization in near-linear time: A martingale approach,” in *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 1539–1554, 2015.
- [21] W. Chen, Y. Wang, and S. Yang, “Efficient influence maximization in social networks,” in *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 199–208, 2009.
- [22] Z. Wen, B. Kveton, and M. Valko, “Online influence maximization under independent cascade model with semi-bandit feedback,” *arXiv preprint arXiv:1605.06593*, 2017.
- [23] S. Vaswani, L. Lakshmanan, M. Schmidt, *et al.*, “Influence maximization with bandits,” *arXiv preprint arXiv:1503.00024*, 2015.
- [24] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, “Combinatorial multi-armed bandit and its extension to probabilistically triggered arms,” *J. Mach. Learn. Res. (JMLR)*, vol. 17, no. 1, pp. 1746–1778, 2016.
- [25] T. Lin, J. Li, and W. Chen, “Stochastic online greedy learning with semi-bandit feedbacks,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 352–360, 2015.
- [26] A. O. Saritac, A. Karakurt, and C. Tekin, “Online contextual influence maximization in social networks,” in *Proc. 54th Annual Allerton Conf. Communication, Control, and Computing (Allerton)*, pp. 1204–1211, 2016. © 2016 IEEE. Reprinted, with permission.

- [27] Y. Bao, X. Wang, Z. Wang, C. Wu, and F. C. M. Lau, “Online influence maximization in non-stationary social networks,” in *24th IEEE/ACM International Symposium on Quality of Service*, pp. 1–6, 2016.
- [28] W. Chen, C. Wang, and Y. Wang, “Scalable influence maximization for prevalent viral marketing in large-scale social networks,” in *Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1029–1038, 2010.
- [29] J. Kim, S. K. Kim, and H. Yu, “Scalable and parallelizable processing of influence maximization for large-scale social networks,” in *IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 266–277, 2013.
- [30] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, “Maximizing social influence in nearly optimal time,” in *Proc. 25th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ’14*, pp. 946–957, 2014.
- [31] S. Lei, S. Maniu, L. Mo, R. Cheng, and P. Senellart, “Online influence maximization,” in *Proc. 21th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 645–654, 2015.
- [32] W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.
- [33] H. Robbins, “Some aspects of the sequential design of experiments,” *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.
- [34] T. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
- [35] 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.
- [36] S. Bubeck and N. Cesa-Bianchi, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

- [37] A. O. Saritac and C. Tekin, “Combinatorial multi-armed bandit problem with probabilistically triggered arms: A case with bounded regret,” in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 111–115, 2017. © 2017 IEEE. Reprinted, with permission.
- [38] Q. Wang and W. Chen, “Tighter regret bounds for influence maximization and other combinatorial semi-bandits with probabilistically triggered arms,” *arXiv preprint arXiv:1703.01610*, 2017.
- [39] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis, “A structured multiarmed bandit problem and the greedy policy,” *IEEE Trans. Autom. Control*, vol. 54, no. 12, pp. 2787–2802, 2009.
- [40] O. Atan, C. Tekin, and M. v. d. Schaar, “Global multi-armed bandits with Hölder continuity,” in *Proc. AISTATS*, pp. 28–36, 2015.
- [41] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvári, “Tight regret bounds for stochastic combinatorial semi-bandits,” in *Proc. AISTATS*, vol. 38, pp. 535–543, 2015.
- [42] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu, “Combinatorial multi-armed bandit with general reward functions,” in *Proc. NIPS*, pp. 1651–1659, 2016.
- [43] R. Agrawal, “Sample mean based index policies by $O(\log n)$ regret for the multi-armed bandit problem,” *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.
- [44] S. Agrawal and N. Goyal, “Analysis of Thompson sampling for the multi-armed bandit problem,” in *Proc. COLT*, pp. 39.1–39.26, 2012.
- [45] E. Kaufmann, N. Korda, and R. Munos, “Thompson sampling: An optimal finite time analysis,” in *Proc. ALT*, pp. 199–213, 2012.
- [46] S. L. Scott, “A modern Bayesian look at the multi-armed bandit,” *Applied Stochastic Models in Business and Industry*, vol. 26, no. 6, pp. 639–658, 2010.
- [47] O. Granmo, “Solving two-armed Bernoulli bandit problems using a Bayesian learning automaton,” *Int. J. Intelligent Computing and Cybernetics*, vol. 3, no. 2, pp. 207–234, 2010.

- [48] T. Graepel, J. Quiñonero Candela, T. Borchert, and R. Herbrich, “Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft’s Bing Search Engine,” in *Proc. ICML*, pp. 13–20, 2010.
- [49] P. Auer, “Using confidence bounds for exploitation-exploration trade-offs,” *J. Mach. Learn. Res. (JMLR)*, vol. 3, pp. 397–422, 2003.
- [50] V. Dani, T. P. Hayes, and S. M. Kakade, “Stochastic linear optimization under bandit feedback,” in *Proc. COLT*, pp. 355–366, 2008.
- [51] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” in *Proc. NIPS*, pp. 2312–2320, 2011.
- [52] A. Durand and C. Gagné, “Thompson sampling for combinatorial bandits and its application to online feature selection,” in *Proc. AAAI-14 Workshop on Sequential Decision-making with Big Data*, 2014.
- [53] S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi, “Thompson sampling for the MNL-bandit,” *arXiv preprint arXiv:1706.00977*, 2017.
- [54] Q. Wang and W. Chen, “Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1161–1171, 2017.
- [55] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A contextual-bandit approach to personalized news article recommendation,” in *Proc. 19th Int. Conf. World Wide Web*, pp. 661–670, 2010.
- [56] L. Qin, S. Chen, and X. Zhu, “Contextual combinatorial bandit and its application on diversified online recommendation,” in *Proc. 2014 SIAM Int. Conf. Data Mining*, pp. 461–469, 2014.
- [57] S. Li, B. Wang, S. Zhang, and W. Chen, “Contextual combinatorial cascading bandits,” in *Proc. 33rd Int. Conf. Machine Learning*, pp. 1245–1253, 2016.
- [58] V. V. Vazirani, *Approximation algorithms*. Springer, 2001.

- [59] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, “How to use expert advice,” *Journal of the ACM (JACM)*, vol. 44, no. 3, pp. 427–485, 1997.
- [60] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.
- [61] J.-Y. Audibert and S. Bubeck, “Regret bounds and minimax policies under partial monitoring,” *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2785–2836, 2010.
- [62] Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori, “Prediction with limited advice and multiarmed bandits with paid observations,” in *Proc. 31st Int. Conf. Machine Learning*, vol. 32, (Beijing, China), pp. 280–287, June 2014.
- [63] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, “An analysis of approximations for maximizing submodular set functions–I,” *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.
- [64] D. Kalman, “A generalized logarithm for exponential-linear equations,” *The College Mathematics Journal*, vol. 32, no. 1, 2001.

Appendix A

Appendix

A.1 Proofs of Theorems Regarding Algorithms for the CMAB with PTAs

A.1.1 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*

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

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 $\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t} := (\hat{\mu}_1^{(\alpha,\beta),t}, \dots, \hat{\mu}_m^{(\alpha,\beta),t})$. 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 $\Delta_t^{(\alpha,\beta)} := \max_{i \in \{1, \dots, m\}} |\mu_i - \hat{\mu}_i^{(\alpha,\beta),t}|$. 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 $\text{gap}(S_t) \leq 0$, 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

$$\begin{aligned} r_{\boldsymbol{\mu}}(S_t) &\geq r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(S_t) - f(\Delta_t^{(\alpha,\beta)}) \\ &\geq \alpha r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(\hat{S}_t^*) - f(\Delta_t^{(\alpha,\beta)}) \\ &\geq \alpha r_{\hat{\boldsymbol{\mu}}^{(\alpha,\beta),t}}(S^*) - f(\Delta_t^{(\alpha,\beta)}) \\ &\geq \alpha r_{\boldsymbol{\mu}}^* - 2f(\Delta_t^{(\alpha,\beta)}) \end{aligned} \tag{A.1}$$

with probability at least β . Next, we show that (A.1) and $S_t \in \{S_{B,1}, \dots, S_{B,n_B}\}$ cannot hold at the same time. $S_t \in \{S_{B,1}, \dots, S_{B,n_B}\}$ 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 $S_t \notin \{S_{B,1}, \dots, S_{B,n_B}\}$ with probability at least β when $\Delta_t^{(\alpha,\beta)} < f^{-1}(\nabla_{\min}/2)$. $\square$

A.1.2 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. 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 $\lim_{t \rightarrow \infty} g(t) = 0$ 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 [64], we obtain

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

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

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

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

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

A.1.3 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 $t \geq t' = 4c^2/\epsilon^2$,

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

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 (A.3) 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 $\lim_{t \rightarrow \infty} g'_0(t) = 0$ 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 [64], we obtain

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

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

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

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 (A.3) hold for all integers $t \geq \max\{4c_0^2/e^2, 4c^2/e^2\}$. $\square$

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 $i \in \{1, \dots, m\}$ implies that $\Delta_{t+1}^\kappa < \delta$ by the triangle inequality. Hence, by the contrapositive of this proposition, we have $\{\Delta_{t+1}^\kappa \geq \delta\} \subseteq \{\bigcup_i \{|\hat{\mu}_i^{t+1} - \mu_i| \geq \delta/2\}\} \cup \{\bigcup_i \{\kappa \sqrt{\frac{3 \ln t}{2T_i^{t+1}}} \geq$

$\delta/2\}\}$. 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.

A.1.4 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

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

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

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

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

$$\begin{aligned}
\mathbb{E} \left[\sum_{t=1}^T \text{IR}_{\boldsymbol{\mu}, \alpha, \beta}^{\text{CUCB-}\kappa}(t) \right] &\leq \nabla_{\max} \left(\lceil t_1 \rceil + \frac{m\pi^2}{3(1-e^{-\delta^2/2})} + \frac{2m}{1-e^{-\delta^2\eta p^*/2}} + \frac{m\pi^2}{6} \right) \\
&\leq \nabla_{\max} \left(\lceil t_1 \rceil + \frac{m\pi^2}{3} \left(1 + \frac{2}{\delta^2} + \frac{1}{2} \right) + 2m \left(1 + \frac{2}{\delta^2\eta p^*} \right) \right). \tag{A.5}
\end{aligned}$$

Since (A.5) holds for any $\eta \in (0, 1)$, the final result is obtained by minimizing

(A.5) 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

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

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

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

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

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T \text{IR}_{\boldsymbol{\mu}, \alpha, \beta}^{\text{CUCB-}\kappa}(t) \right] &\leq \nabla_{\max} \left(\lceil t' \rceil + \frac{m\pi^2}{3(1-e^{-2\delta^2})} + \frac{2m}{1-e^{-2\delta^2\eta p^*}} \right) \\ &\leq \nabla_{\max} \left(\lceil t' \rceil + \frac{m\pi^2}{3} \left(1 + \frac{1}{2\delta^2} \right) + 2m \left(1 + \frac{1}{2\delta^2\eta p^*} \right) \right). \end{aligned} \quad (\text{A.6})$$

Finally, we minimize (A.6) over $\eta \in (0, 1)$ via the procedure we follow for minimizing (A.5).

A.1.5 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 $\text{IR}_{\boldsymbol{\mu}, \alpha, \beta}^{\text{CUCB-}\kappa}(t)$ can be bounded by using Δ_t^0 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:*

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

Proof. Taking expectation of both sides of the equation (A.1) in Lemma 1, we obtain the desired result. $\square$

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

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

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

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

Proof. We have

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

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

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

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

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

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

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

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

$2m + 1 \leq 3m$, we have

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

□

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

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

Now, we can bound the regret by bounding the sum of expected instantaneous

regrets throughout the time horizon:

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

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 $\omega \in (0, 1]$:

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

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

A.1.6 Proof of Theorem 5

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

Lemma 7. (Lemma 9 in [44]) $F_{n,p}^{\text{Binom}}(np - n\delta) \leq e^{-2n\delta^2}$, $1 - F_{n,p}^{\text{Binom}}(np + n\delta) \leq e^{-2n\delta^2}$ and $1 - F_{n+1,p}^{\text{Binom}}(np + n\delta) \leq e^{4\delta - 2n\delta^2}$ for all $n \in \mathbb{N}_+$, $p \in [0, 1]$, $\delta \geq 0$.

Fact 1. $F_{n+1,p}^{\text{Binom}}(i) \geq F_{n,p}^{\text{Binom}}(i-1)$ 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 [44])

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

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:

$$\begin{aligned} & \Pr(|\nu_i^{t+1} - \mu_i| \geq \delta) \\ &= \Pr(|\nu_i^{t+1} - \mu_i| \geq \delta, \Delta_{t+1}^0 \geq \delta/2) + \Pr(|\nu_i^{t+1} - \mu_i| \geq \delta, \Delta_{t+1}^0 < \delta/2) \\ &\leq \Pr(\Delta_{t+1}^0 \geq \delta/2) + \Pr(|\nu_i^{t+1} - \mu_i| \geq \delta, \Delta_{t+1}^0 < \delta/2). \end{aligned} \quad (\text{A.7})$$

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

$$\Pr(\Delta_{t+1}^0 < \delta/2, |\nu_i^{t+1} - \mu_i| \geq \delta) \leq \Pr(|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2). \quad (\text{A.8})$$

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

$$\begin{aligned} & \Pr(|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2) \\ &= \sum_{j=0}^t \Pr(|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2 | T_i^{t+1} = j) \Pr(T_i^{t+1} = j) \\ &\leq \sum_{j=0}^t \Pr(\nu_i^{t+1} - \hat{\mu}_i^{t+1} \geq \delta/2 | T_i^{t+1} = j) \Pr(T_i^{t+1} = j) \\ &\quad + \sum_{j=0}^t \Pr(\nu_i^{t+1} - \hat{\mu}_i^{t+1} \leq -\delta/2 | T_i^{t+1} = j) \Pr(T_i^{t+1} = j). \end{aligned} \quad (\text{A.9})$$

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

$$\begin{aligned}
\Pr(\nu_i^{t+1} - \hat{\mu}_i^{t+1} \geq \delta/2 | T_i^{t+1} = j) &= \mathbb{E}[1 - F_{1+s_i^{t+1}, j-s_i^{t+1}+1}^{\text{Beta}}(\hat{\mu}_i^{t+1} + \delta/2) | T_i^{t+1} = j] \\
&= \mathbb{E}[F_{j+1, \hat{\mu}_i^{t+1} + \delta/2}^{\text{Binom}}(s_i^{t+1}) | T_i^{t+1} = j] \\
&\leq \mathbb{E}[F_{j, \hat{\mu}_i^{t+1} + \delta/2}^{\text{Binom}}(s_i^{t+1}) | T_i^{t+1} = j] \\
&= \mathbb{E}[F_{j, \hat{\mu}_i^{t+1} + \delta/2}^{\text{Binom}}(j\hat{\mu}_i^{t+1}) | T_i^{t+1} = j] \\
&\leq e^{-\delta^2 j/2}.
\end{aligned} \tag{A.10}$$

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

$$\begin{aligned}
\Pr(\nu_i^{t+1} - \hat{\mu}_i^{t+1} \leq -\delta/2 | T_i^{t+1} = j) &= \mathbb{E}[F_{1+s_i^{t+1}, j-s_i^{t+1}+1}^{\text{Beta}}(\hat{\mu}_i^{t+1} - \delta/2) | T_i^{t+1} = j] \\
&= \mathbb{E}[1 - F_{j+1, \hat{\mu}_i^{t+1} - \delta/2}^{\text{Binom}}(s_i^{t+1}) | T_i^{t+1} = j] \\
&= \mathbb{E}[1 - F_{j+1, \hat{\mu}_i^{t+1} - \delta/2}^{\text{Binom}}(j\hat{\mu}_i^{t+1}) | T_i^{t+1} = j] \\
&\leq e^{2\delta - \delta^2 j/2}.
\end{aligned} \tag{A.11}$$

Using (A.10) and (A.11) in (A.9), we obtain

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

For $0 \leq j < t^* = \eta p^* t$, we have $\Pr(T_i^{t+1} = j) \leq \Pr(T_i^{t+1} \leq t^*) \leq 1/t^2$ 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

$$\begin{aligned}
& \frac{\Pr(|\nu_i^{t+1} - \mu_i| \geq \delta, \Delta_{t+1}^0 < \delta/2)}{e^{2\delta} + 1} \\
& \leq \frac{\Pr(|\nu_i^{t+1} - \hat{\mu}_i^{t+1}| \geq \delta/2)}{e^{2\delta} + 1} \\
& \leq \sum_{j=0}^{\lfloor t^* \rfloor} e^{-\delta^2 j/2} \Pr(T_i^{t+1} = j) + \sum_{j=\lceil t^* \rceil}^t e^{-\delta^2 j/2} \Pr(T_i^{t+1} = j) \\
& \leq \frac{1}{t^2} \sum_{j=0}^{\lfloor t^* \rfloor} e^{-\delta^2 j/2} + e^{-\delta^2 \lceil t^* \rceil/2} \sum_{j=\lceil t^* \rceil}^t \Pr(T_i^{t+1} = j) \\
& \leq \frac{1}{t^2(1 - e^{-\delta^2/2})} + e^{-\delta^2 \lceil t^* \rceil/2} \\
& \leq \frac{1}{t^2(1 - e^{-\delta^2/2})} + e^{-\delta^2 \eta p^* t/2}. \tag{A.12}
\end{aligned}$$

Using the result of Lemma 4 and (A.12) in (A.7), and recalling the inequality in (A.8), we obtain

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

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

Finally by using the union bound, we obtain

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

for $t \geq t'$.

A.1.7 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

$\Delta_t^\nu = \max_i |\nu_i^t - \mu_i|$. By Lemma 1, we have

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

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

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T \text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) \right] &= \mathbb{E} \left[\sum_{t=1}^{\lceil t' \rceil} \text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) \right] + \mathbb{E} \left[\sum_{t=\lceil t' \rceil+1}^T \text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) \right] \\ &\leq \nabla_{\max} \lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^T \mathbb{E}[\text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) | \Delta_t^\nu < \delta] \Pr(\Delta_t^\nu < \delta) \\ &\quad + \sum_{t=\lceil t' \rceil+1}^T \mathbb{E}[\text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) | \Delta_t^\nu \geq \delta] \Pr(\Delta_t^\nu \geq \delta) \\ &= \nabla_{\max} \lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^T \mathbb{E}[\text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) | \mathcal{N}_t] \Pr(\Delta_t^\nu < \delta) \\ &\quad + \sum_{t=\lceil t' \rceil+1}^T \mathbb{E}[\text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) | \Delta_t^\nu \geq \delta] \Pr(\Delta_t^\nu \geq \delta) \\ &\leq \nabla_{\max} \lceil t' \rceil + \sum_{t=\lceil t' \rceil+1}^T \mathbb{E}[\text{IR}_{\mu, \alpha, \beta}^{\text{CTS}}(t) | \Delta_t^\nu \geq \delta] \\ &\quad \times \left(\frac{(3 + e^{2\delta})m}{(t-1)^2(1 - e^{-\delta^2/2})} + (3 + e^{2\delta})me^{-\delta^2 \eta p^* (t-1)/2} \right) \\ &\leq \nabla_{\max} \lceil t' \rceil + \nabla_{\max} \sum_{t=1}^{\infty} \left(\frac{(3 + e^{2\delta})m}{t^2(1 - e^{-\delta^2/2})} + (3 + e^{2\delta})me^{-\delta^2 \eta p^* t/2} \right) \\ &\leq \nabla_{\max} \left(\lceil t' \rceil + \frac{(3 + e^{2\delta})m\pi^2}{6(1 - e^{-\delta^2/2})} + \frac{(3 + e^{2\delta})me^{-\delta^2 \eta p^*/2}}{1 - e^{-\delta^2 \eta p^*/2}} \right) \\ &\leq \nabla_{\max} \left(\lceil t' \rceil + \frac{(3 + e^{2\delta})m\pi^2}{6(1 - e^{-\delta^2/2})} + \frac{(3 + e^{2\delta})m}{1 - e^{-\delta^2 \eta p^*/2}} \right). \end{aligned}$$

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

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

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

A.1.8 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 $\text{IR}_{\boldsymbol{\mu}, \alpha, \beta}^{\text{CTS}}(t)$ can be bounded by using Δ_t^ν 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:*

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

Proof. Taking expectation of both sides of the equation (A.1) in Lemma 1, we obtain the desired result. $\square$

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

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

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

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

Proof. We have

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

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

$$\begin{aligned}\mathbb{E}[\Delta_{t+1}^\nu] &\leq \int_0^{t^{-\theta}} \Pr(\Delta_{t+1}^\nu \geq \delta) d\delta + \int_{t^{-\theta}}^1 \left(\frac{(3+e^{2\delta})m}{t^2(1-e^{-\delta^2/2})} + (3+e^{2\delta})me^{-\delta^2\eta p^* t/2} \right) d\delta \\ &\leq t^{-\theta} + \int_{t^{-\theta}}^1 \left(\frac{(3+e^2)m}{t^2(1-e^{-\delta^2/2})} + (3+e^2)me^{-\delta^2\eta p^* t/2} \right) d\delta\end{aligned}$$

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

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

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

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

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

$$\mathbb{E}[\Delta_{t+1}^\nu] \leq (3 + e^2)m\sqrt{\frac{\pi}{At}} + (2(3 + e^2)m + 1)t^{-1} \quad (\text{A.13})$$

$$\begin{aligned} &\leq (3 + e^2)mt^{-1/2}\sqrt{\frac{\pi}{A}} + 3(3 + e^2)mt^{-1/2} \\ &\leq (3 + e^2)mt^{-1/2}\left(\sqrt{\frac{\pi}{A}} + 3\right). \end{aligned} \quad (\text{A.14})$$

□

Next, using Lemma 9, we show that the gap-independent regret for CTS is at most $O(T^{1-\omega/2})$ 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, (A.14) and Jensen's inequality, we have

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

Now, we can bound the regret by bounding the sum of expected instantaneous

regrets throughout the time horizon:

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

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 $\omega \in (0, 1]$:

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

A.2 Proofs of Theorems Regarding Algorithms for OCIMP

A.2.1 Proof Of Theorem 9

By definition of the (α, β) -approximation oracle, we have $\hat{\sigma}(x, \hat{S}^{(\alpha, \beta)}(x)) \geq \alpha \times \hat{\sigma}(x, \hat{S}^*(x))$ with probability at least β . Theorem 8 implies that for any seed set S , $|\hat{\sigma}(x, S) - \sigma(x, S)| \leq mn\Delta$. Using the results above, we obtain

$$\begin{aligned}
\sigma(x, \hat{S}^{(\alpha, \beta)}(x)) &\geq \hat{\sigma}(x, \hat{S}^{(\alpha, \beta)}(x)) - mn\Delta \\
&\geq \alpha \hat{\sigma}(x, \hat{S}^*(x)) - mn\Delta \\
&\geq \alpha \hat{\sigma}(x, S^*(x)) - mn\Delta \\
&\geq \alpha (\sigma(x, S^*(x)) - mn\Delta) - mn\Delta \\
&= \alpha \sigma(x, S^*(x)) - (1 + \alpha)mn\Delta
\end{aligned}$$

with probability at least β . Since $\sigma(x, \hat{S}^{(\alpha, \beta)}(x))$ is non-negative, we obtain the following bound by taking the expectation:

$$\mathbb{E}[\sigma(x, \hat{S}^{(\alpha, \beta)}(x))] \geq \alpha\beta \times \sigma(x, S^*(x)) - \beta(1 + \alpha)mn\Delta.$$

A.2.2 Proof of Theorems 10 and 11

For $Q \in \mathcal{Q}$ let $\bar{p}_{i,j}^Q := \sup_{x \in Q} p_{i,j}^x$ and $\underline{p}_{i,j}^Q := \inf_{x \in Q} p_{i,j}^x$. Consider an algorithm π (COIN-CO-EL or COIN-CO-NL) with partitioning parameter $q_T = \lceil T^z \rceil$ and control function $D(t) = (c+1)^{\eta t^\gamma}$, where $0 < \gamma, z < 1$ and $\eta < 0$. For any sequence of context arrivals $\{x_t\}_{t=1}^T$, let $\mathcal{T}_T^{s,\pi}$ be the set of epochs by epoch T in which algorithm π exploits and $\mathcal{T}_T^{o,\pi}$ be the set of epochs by epoch T in which π explores. Since the activation attempts are random variables, $\mathcal{T}_T^{s,\pi}$ and $\mathcal{T}_T^{o,\pi}$ are random sets for which $\mathcal{T}_T^{s,\pi} \cup \mathcal{T}_T^{o,\pi} = \{1, \dots, T\}$ with probability 1. By the definition of the exploration and exploitation phases of COIN-CO-EL and COIN-CO-NL, we have for any $t \in \mathcal{T}_T^{s,\pi}$

$$f_{i,j}^{Q_t}(t) + s_{i,j}^{Q_t}(t) \geq D(t) \quad \forall (i, j) \in E. \quad (\text{A.15})$$

The *simple* (α, β) -regret of algorithm π for epoch t is defined as

$$r_\pi^{(\alpha, \beta)}(t) := \alpha\beta \times \sigma(x_t, S^*(x_t)) - \sigma(x_t, S_t).$$

Let

$$R_\pi^s(T) := \sum_{t \in \mathcal{T}_T^{s,\pi}} r_\pi^{(\alpha, \beta)}(t)$$

be the regret incurred over epochs in which algorithm π exploits,

$$R_\pi^o(T) := \sum_{t \in \mathcal{T}_T^{o,\pi}} r_\pi^{(\alpha, \beta)}(t)$$

be the regret (except the cost of observing the influence outcomes or activated nodes) incurred over epochs in which algorithm π explores, and

$$R_\pi^c(T) = c \times \sum_{t=1}^T B_t$$

be the regret due to observing the influence outcomes(or activated nodes) over epochs in which algorithm π explores. Based on the above definitions, the regret can be decomposed as follows:

$$\mathbb{E}[R_\pi^{(\alpha,\beta)}(T)] = \mathbb{E}[R_\pi^s(T)] + \mathbb{E}[R_\pi^o(T)] + \mathbb{E}[R_\pi^c(T)].$$

The proof proceeds with bounding each of the terms given above. First, we bound $R_\pi^o(T)$ for COIN-CO-EL and COIN-CO-NL .

Lemma 10. *When $\pi = \text{COIN-CO-EL}$ runs with control function $D(t) = (c+1)^\eta t^\gamma$ and partitioning parameter $q_T = \lceil T^z \rceil$, where $0 < \gamma, z < 1$ and $\eta < 0$, we have*

$$R_\pi^o(T) \leq \alpha\beta(n-k) \left(\frac{m}{k} + 1 \right) \lceil T^z \rceil^d \lceil (c+1)^\eta T^\gamma \rceil$$

with probability 1.

Proof. At each epoch when COIN-CO-EL explores, in the worst case, COIN-CO-EL will fail to influence the remaining $(n-k)$ nodes and the omnipotent oracle will influence all the remaining $(n-k)$ nodes. Hence, we have $r_\pi^{(\alpha,\beta)}(t) \leq \alpha\beta(n-k)$ for each $t \in \mathcal{T}_T^{o,\pi}$, which implies that

$$R_\pi^o(T) \leq \alpha\beta(n-k)|\mathcal{T}_T^{o,\pi}| \text{ with probability 1.} \quad (\text{A.16})$$

Next, we will bound $|\mathcal{T}_T^{o,\pi}|$. Let $\mathcal{T}_T^L$ denote the set of exploration epochs of COIN-CO-EL where $|Y_{Q_i}(t)| < k$ and $\mathcal{T}_T^H$ denote the set of exploration epochs where $|Y_{Q_i}(t)| \geq k$. We have $\mathcal{T}_T^{o,\pi} = \mathcal{T}_T^H \cup \mathcal{T}_T^L$.

Firstly, we bound $|\mathcal{T}_T^L|$. Note that for each $Q \in \mathcal{Q}$, there will be at most

$\lceil D(T) \rceil$ many epochs where $Q_t = Q$ and $|Y_{Q_t}(t)| < k$. Therefore,

$$|\mathcal{T}_T^L| < q_T^d \lceil D(T) \rceil \text{ with probability 1.} \quad (\text{A.17})$$

Secondly, we bound $|\mathcal{T}_T^H|$. Let $u(t) := |F_t \cap Y_{Q_t}(t)|$. Note that given T , we have a total of mq_T^d many context set-edge pairs. Hence, the total number of explorations can be at most $mq_T^d \lceil D(T) \rceil$. Thus

$$\sum_{t \in \mathcal{T}_T^H} u(t) \leq mq_T^d \lceil D(T) \rceil \text{ with probability 1.} \quad (\text{A.18})$$

From the definition of $\mathcal{T}_T^H$, we know that $u(t) \geq k$ for all $t \in \mathcal{T}_T^H$. Using this together with (A.18), we obtain $k|\mathcal{T}_T^H| \leq mq_T^d \lceil D(T) \rceil$, and hence,

$$|\mathcal{T}_T^H| \leq \frac{mq_T^d \lceil D(T) \rceil}{k} \text{ with probability 1.} \quad (\text{A.19})$$

Hence, by summing (A.17) and (A.19), we obtain

$$|\mathcal{T}_T^{o,\pi}| \leq \left(\frac{m}{k} + 1\right) \lceil T^z \rceil^d \lceil (c+1)^\eta T^\gamma \rceil \text{ with probability 1.} \quad (\text{A.20})$$

The result follows by plugging the bound in (A.20) into (A.16). $\square$

Lemma 11. *When $\pi = \text{COIN-CO-NL}$ runs with control function $D(t) = (c+1)^\eta t^\gamma$ and partitioning parameter $q_T = \lceil T^z \rceil$, where $0 < \gamma, z < 1$ and $\eta < 0$, we have*

$$R_\pi^o(T) \leq \alpha\beta(n-k)m \lceil T^z \rceil^d \lceil (c+1)^\eta T^\gamma \rceil$$

with probability 1.

Proof. At each epoch when COIN-CO-NL explores, in the worst case, COIN-CO-NL will fail to influence the remaining $(n-k)$ nodes and the omnipotent oracle will influence all the remaining $(n-k)$ nodes. Hence, we have $r_\pi^{(\alpha,\beta)}(t) \leq \alpha\beta(n-k)$ for each $t \in \mathcal{T}_T^{o,\pi}$, which implies that

$$R_\pi^o(T) \leq \alpha\beta(n-k)|\mathcal{T}_T^{o,\pi}| \text{ with probability 1.} \quad (\text{A.21})$$

Next, we will bound $|\mathcal{T}_T^{o,\pi}|$. Note that COIN-CO-NL explores (infers the influence outcome on) at least one edge in each exploration epoch, and each context partition-edge pair will be explored at most $q_T^d \lceil D(T) \rceil$ times. Hence, we have

$$|\mathcal{T}_T^{o,\pi}| \leq m q_T^d \lceil D(T) \rceil \text{ with probability 1.}$$

which gives the desired result when used together with (A.21). $\square$

Next, we bound the regret due to costly influence outcome (node activation) observations for COIN-CO-EL (COIN-CO-NL).

Lemma 12. *When COIN-CO-EL or COIN-CO-NL runs with control function $D(t) = (c+1)^\eta t^\gamma$ and partitioning parameter $q_T = \lceil T^z \rceil$, where $0 < \gamma, z < 1$ and $\eta < 0$, we have*

$$R_\pi^c(T) \leq cm \lceil T^z \rceil^d \lceil (c+1)^\eta T^\gamma \rceil \text{ with probability 1.}$$

Proof. Since COIN-CO-EL and COIN-CO-NL both keep an influence probability estimate for each edge $(i, j) \in E$ for each $Q \in \mathcal{Q}$, they keep $m q_T^d$ parameters to represent the influence probability estimates. At each exploration epoch t , they observe the influence outcomes on the edges that correspond to a subset of the edges which are explored less than $D(t)$ times (COIN-CO-EL) or node activations for a subset of nodes that are adjacent to a subset of the under-explored edges (COIN-CO-NL). Thus, by the end of epoch T , the influence outcome on an edge (or a node) is observed at most $\lceil D(T) \rceil q_T^d$ times. Therefore, the total number of costly observations is at most $m \lceil D(T) \rceil q_T^d$ and each of these observations costs c . $\square$

Next, we bound the exploitation regret. For this, we first propose the following lemma, which bounds $r_\pi^{(\alpha,\beta)}(t)$ when COIN-CO-EL or COIN-CO-NL exploits at epoch t .

Lemma 13. *When $\pi = \text{COIN-CO-EL}$ or $\pi = \text{COIN-CO-NL}$ runs with control function $D(t) = (c+1)^\eta t^\gamma$ and partitioning parameter $q_T = \lceil T^z \rceil$, where $0 <$*

$\gamma, z < 1$ and $\eta < 0$, we have

$$\begin{aligned}\mathbb{E}[r_\pi^{(\alpha, \beta)}(t) | t \in \mathcal{T}_T^{s, \pi}] &\leq \beta(1 + \alpha)mnLd^{\theta/2}q_T^{-\theta} \\ &\quad + \frac{\beta(1 + \alpha)\pi m^2 n t^{-\gamma/2} (c + 1)^{-\eta/2}}{\sqrt{2}}.\end{aligned}$$

Proof. In the analysis below, we consider $t \in \mathcal{T}_T^{s, \pi}$, hence all the expectations are conditioned on this event. Let $\Delta_t := \Delta_{x_t}(\mathbf{p}, \hat{\mathbf{p}}_t)$. By Theorem 9, we have

$$\begin{aligned}\mathbb{E}[r_\pi^{(\alpha, \beta)}(t)] &= \alpha\beta \times \sigma(x_t, S^*(x_t)) - \mathbb{E}[\sigma(x_t, S_t)] \\ &\leq \beta(1 + \alpha)mn\mathbb{E}[\Delta_t].\end{aligned}$$

Since $\Delta_t \in [0, 1]$, we have

$$\mathbb{E}[r_\pi^{(\alpha, \beta)}(t)] \leq \beta(1 + \alpha)mn \int_0^1 \Pr(\Delta_t \geq y) dy. \quad (\text{A.22})$$

Note that

$$\begin{aligned}\{\Delta_t \geq y\} &= \left\{ \max_{(i, j) \in E} (|\hat{p}_{i, j}^{Q_t}(t) - p_{i, j}^{x_t}| \geq y) \right\} \\ &= \bigcup_{(i, j) \in E} \left\{ |\hat{p}_{i, j}^{Q_t}(t) - p_{i, j}^{x_t}| \geq y \right\} \\ &= \bigcup_{(i, j) \in E} \{ \hat{p}_{i, j}^{Q_t}(t) - p_{i, j}^{x_t} \leq -y \} \cup \bigcup_{(i, j) \in E} \{ \hat{p}_{i, j}^{Q_t}(t) - p_{i, j}^{x_t} \geq y \} \\ &\subset \bigcup_{(i, j) \in E} \{ \hat{p}_{i, j}^{Q_t}(t) - \bar{p}_{i, j}^{Q_t} \leq -y \} \cup \bigcup_{(i, j) \in E} \{ \hat{p}_{i, j}^{Q_t}(t) - \underline{p}_{i, j}^{Q_t} \geq y \}.\end{aligned}$$

Hence, by the union bound, we get

$$\begin{aligned}\Pr(\Delta_t \geq y) &\leq \sum_{(i, j) \in E} \Pr(\hat{p}_{i, j}^{Q_t}(t) - \bar{p}_{i, j}^{Q_t} \leq -y) \\ &\quad + \sum_{(i, j) \in E} \Pr(\hat{p}_{i, j}^{Q_t}(t) - \underline{p}_{i, j}^{Q_t} \geq y).\end{aligned}$$

By Assumption 3, we have $\bar{p}_{i, j}^{Q_t} - \underline{p}_{i, j}^{Q_t} \leq Ld^{\theta/2}q_T^{-\theta}$ for all $(i, j) \in E$ and $Q_t \in \mathcal{Q}$. Hence, we have $\bar{p}_{i, j}^{Q_t} \leq \mathbb{E}[\hat{p}_{i, j}^{Q_t}(t)] + Ld^{\theta/2}q_T^{-\theta}$ and $\underline{p}_{i, j}^{Q_t} \geq \mathbb{E}[\hat{p}_{i, j}^{Q_t}(t)] - Ld^{\theta/2}q_T^{-\theta}$ for all

$(i, j) \in E$ and $Q_t \in \mathcal{Q}$. Using the fact above, we obtain

$$\begin{aligned} & \Pr(\hat{p}_{i,j}^{Q_t}(t) - \bar{p}_{i,j}^{Q_t} \leq -y) \\ & \leq \Pr(\hat{p}_{i,j}^{Q_t}(t) - \mathbb{E}[\hat{p}_{i,j}^{Q_t}(t)] \leq Ld^{\theta/2}q_T^{-\theta} - y) \end{aligned}$$

and

$$\begin{aligned} & \Pr(\hat{p}_{i,j}^{Q_t}(t) - \underline{p}_{i,j}^{Q_t} \geq y) \\ & \leq \Pr(\hat{p}_{i,j}^{Q_t}(t) - \mathbb{E}[\hat{p}_{i,j}^{Q_t}(t)] \geq y - Ld^{\theta/2}q_T^{-\theta}). \end{aligned}$$

Since (A.15) holds for $t \in \mathcal{T}_T^{s,\pi}$, by using the above inequalities together with Hoeffding's inequality, we obtain the following for $y \geq Ld^{\theta/2}q_T^{-\theta}$:

$$\sum_{(i,j) \in E} \Pr(\hat{p}_{i,j}^{Q_t}(t) - \underline{p}_{i,j}^{Q_t} \geq y) \leq me^{-2(y-Ld^{\theta/2}q_T^{-\theta})^2(c+1)\eta t^\gamma} \quad (\text{A.23})$$

$$\sum_{(i,j) \in E} \Pr(\hat{p}_{i,j}^{Q_t}(t) - \bar{p}_{i,j}^{Q_t} \leq -y) \leq me^{-2(y-Ld^{\theta/2}q_T^{-\theta})^2(c+1)\eta t^\gamma}. \quad (\text{A.24})$$

In order to bound (A.22), we will separate the integral into two parts. For $0 \leq y < Ld^{\theta/2}q_T^{-\theta}$, we have $\Pr(\Delta_t \geq y) \leq 1$. For $Ld^{\theta/2}q_T^{-\theta} \leq y \leq 1$ by (A.23) and (A.24), we have $\Pr(\Delta_t \geq y) \leq 2me^{-2(y-Ld^{\theta/2}q_T^{-\theta})^2(c+1)\eta t^\gamma}$. Hence,

$$\begin{aligned} & \int_0^1 \Pr(\Delta_t \geq y) dy \\ & = \int_0^{Ld^{\theta/2}q_T^{-\theta}} 1 dy + \int_{Ld^{\theta/2}q_T^{-\theta}}^1 2me^{-2(y-Ld^{\theta/2}q_T^{-\theta})^2(c+1)\eta t^\gamma} dy \\ & \leq Ld^{\theta/2}q_T^{-\theta} \\ & \quad + 2m \int_{Ld^{\theta/2}q_T^{-\theta}}^1 \frac{dy}{1 + 2(y - Ld^{\theta/2}q_T^{-\theta})^2(c+1)\eta t^\gamma} \\ & = Ld^{\theta/2}q_T^{-\theta} + 2m \frac{(c+1)^{-\eta/2}t^{-\gamma/2}}{\sqrt{2}} \\ & \quad \times (\arctan(\sqrt{2}t^{\gamma/2}(c+1)^{\eta/2}(1 - Ld^{\theta/2}q_T^{-\theta})) \\ & \leq Ld^{\theta/2}q_T^{-\theta} + \frac{m(c+1)^{-\eta/2}t^{-\gamma/2}\pi}{\sqrt{2}} \end{aligned} \quad (\text{A.25})$$

since $e^{-y} \leq \frac{1}{1+y}$ for all $y \geq 0$ and that $\arctan(z) \leq \frac{\pi}{2}$ for all $z \in \mathbb{R}$. The result is obtained by substituting (A.25) in (A.22). $\square$

The next lemma uses Lemma 13 to bound $\mathbb{E}[R_\pi^s(T)]$.

Lemma 14. *When $\pi = \text{COIN-CO-EL}$ or $\pi = \text{COIN-CO-NL}$ runs with control function $D(t) = (c+1)^\eta t^\gamma$ and partitioning parameter $q_T = \lceil T^z \rceil$, where $0 < \gamma, z < 1$, $\eta < 0$, we have*

$$\begin{aligned} \mathbb{E}[R_\pi^s(T)] &\leq \beta(1+\alpha)mnLd^{\theta/2}T^{1-\theta z} \\ &\quad + \frac{\beta(1+\alpha)\pi m^2 n(c+1)^{-\eta/2}}{\sqrt{2}} \times \frac{T^{1-\gamma/2} - \gamma/2}{(1-\gamma/2)}. \end{aligned}$$

Proof. We utilize the following inequalities in the proof: $|\mathcal{T}_T^{s,\pi}| \leq T$ with probability 1 and $\sum_{t=1}^T t^{-x} \leq \frac{T^{1-x}-x}{(1-x)} \forall x \in (0, 1)$. For any realization of $\mathcal{T}_T^{s,\pi}$ denoted by $\mathcal{T} \subset \{1, \dots, T\}$ we have

$$\begin{aligned} \mathbb{E}[R_\pi^s(T) | \mathcal{T}_T^{s,\pi} = \mathcal{T}] &= \sum_{t \in \mathcal{T}} \mathbb{E}[r_\pi^{(\alpha,\beta)}(t) | t \in \mathcal{T}] \\ &\leq \beta(1+\alpha) \\ &\quad \times \sum_{t \in \mathcal{T}} \left(mnLd^{\theta/2} \lceil T^z \rceil^{-\theta} + \frac{\pi m^2 n t^{-\gamma/2} (c+1)^{-\eta/2}}{\sqrt{2}} \right) \\ &\leq \beta(1+\alpha) \\ &\quad \times \sum_{t=1}^T \left(mnLd^{\theta/2} \lceil T^z \rceil^{-\theta} + \frac{\pi m^2 n t^{-\gamma/2} (c+1)^{-\eta/2}}{\sqrt{2}} \right) \\ &\leq \beta(1+\alpha) \\ &\quad \times \sum_{t=1}^T \left(mnLd^{\theta/2} T^{-z\theta} + \frac{\pi m^2 n t^{-\gamma/2} (c+1)^{-\eta/2}}{\sqrt{2}} \right) \\ &\leq \beta(1+\alpha)mnLd^{\theta/2}T^{1-\theta z} \\ &\quad + \frac{\beta(1+\alpha)\pi m^2 n(c+1)^{-\eta/2}}{\sqrt{2}} \times \frac{T^{1-\gamma/2} - \gamma/2}{(1-\gamma/2)}. \end{aligned}$$

$\square$

By summing the results of Lemmas 10, 12 and 14, we obtain for $\pi = \text{COIN-CO-EL}$:

$$\begin{aligned}
\mathbb{E}[R_{\pi}^{(\alpha, \beta)}(T)] &\leq \alpha\beta(n-k) \left(\frac{m}{k} + 1\right) \lceil T^z \rceil^d \lceil (c+1)^{\eta} T^{\gamma} \rceil \\
&+ \beta(1+\alpha)mnLd^{\theta/2}T^{1-\theta z} \\
&+ cm \lceil (c+1)^{\eta} T^{\gamma} \rceil \lceil T^z \rceil^d \\
&+ \frac{\beta(1+\alpha)\pi m^2 n (c+1)^{-\eta/2}}{\sqrt{2}} \times \frac{T^{1-\gamma/2} - \gamma/2}{(1-\gamma/2)}. \tag{A.26}
\end{aligned}$$

Similarly, by summing the results of Lemmas 11, 12 and 14, we obtain for $\pi = \text{COIN-CO-NL}$:

$$\begin{aligned}
\mathbb{E}[R_{\pi}^{(\alpha, \beta)}(T)] &\leq \alpha\beta(n-k)m \lceil T^z \rceil^d \lceil (c+1)^{\eta} T^{\gamma} \rceil \\
&+ \beta(1+\alpha)mnLd^{\theta/2}T^{1-\theta z} \\
&+ cm \lceil (c+1)^{\eta} T^{\gamma} \rceil \lceil T^z \rceil^d \\
&+ \frac{\beta(1+\alpha)\pi m^2 n (c+1)^{-\eta/2}}{\sqrt{2}} \times \frac{T^{1-\gamma/2} - \gamma/2}{(1-\gamma/2)}. \tag{A.27}
\end{aligned}$$

Finally, we calculate the optimal values of the parameters, which minimize the regrets given in (A.26) and (A.27). We observe that the terms in the regret bounds with the highest time orders are $O(T^{zd+\gamma})$, $O(T^{1-\gamma/2})$ and $O(T^{1-\theta z})$. Hence, the optimal z and γ should minimize $\max\{zd+\gamma, 1-\gamma/2, 1-\theta z\}$. This is achieved by setting $z = 1/(3\theta+d)$ and $\gamma = 2\theta/(3\theta+d)$. In addition, we also minimize the order of the cost in the regret bound, given the optimal z and γ values for the time order of the regret. For this, we look at the regret terms whose time orders are $T^{(2\theta+d)/(3\theta+d)}$. The cost order of these terms are $O((c+1)^{1+\eta})$ and $O((c+1)^{-\eta/2})$. Hence, to balance these terms, we set $\eta = -2/3$.

A.2.3 Proof of Theorem 12

Our proof is built on the proof of Theorem 4 in [62]. In the proof, we assume that the learner is deterministic and makes a fixed number of observations, denoted by OT , by epoch T . It is well known that lower bound results for deterministic learners apply for stochastic learners as well [60].

First Step: Influence Graph and a Regret Lower Bound

We define a specific directed graph $\bar{G}(\bar{V}, \bar{E})$ with the following properties. Assume that n is even and $m = n/2$. Let $\bar{V}_0 := \{v_0^1, \dots, v_0^{n/2}\}$, $\bar{V}_1 := \{v_1^1, \dots, v_1^{n/2}\}$, $\bar{E} := \{(v_0^1, v_1^1), (v_0^2, v_1^2), \dots, (v_0^m, v_1^m)\}$ and $\bar{V} := \bar{V}_0 \cup \bar{V}_1$. Since the nodes in $\bar{V}_1$ cannot influence any other node, any sensible policy will only select nodes from $\bar{V}_0$ as the seed set. Let $A = \binom{m}{k}$ denote the cardinality of the action set $\mathcal{M}$. We index the actions in a way that V_i denotes the i th action, and hence, $\mathcal{M} := \{V_1, \dots, V_A\}$. Due to the fact that each node in $\bar{V}_1$ has a single parent, in this setting edge-level and node-level feedbacks are equivalent. Thus, in the rest of the proof, we focus on only edge-level feedback. We assume that the influence probabilities are independent of context, we simplify the notation and use $\sigma(S)$ and S^* to denote the expected influence spread of action S and an optimal action, respectively.

We define $m + 1$ problem instances on $\bar{G}(\bar{V}, \bar{E})$, indexed by $\{0, 1, \dots, m\}$. For $h > 0$, in the h th problem instance, we set the influence probabilities on the edges as $p_{v_0^h, v_1^h} = \frac{1+\epsilon}{2}$, and $p_i = \frac{1-\epsilon}{2}$ for $i \in \bar{E} - (v_0^h, v_1^h)$. Moreover, in the 0th problem instance, we set $p_i = \frac{1-\epsilon}{2}$ for all $i \in \bar{E}$. Note that with the above construction, for any problem instance $h > 0$, there exists $\binom{m-1}{k-1}$ actions with expected influence spread $(k-1)\frac{1-\epsilon}{2} + \frac{1+\epsilon}{2}$, and $\binom{m}{k} - \binom{m-1}{k-1} = \binom{m-1}{k}$ actions with expected influence spread $k\frac{1-\epsilon}{2}$.

Let $\hat{q} = \{\hat{q}_1, \dots, \hat{q}_A\}$ be the marginal distribution over the actions selected by the learner over time horizon T , where $\hat{q}_i = \frac{1}{T} \sum_{t=1}^T \mathbf{1}_{\{S_t=V_i\}}$ and $\mathbf{1}_{\{\cdot\}}$ denotes the indicator variable which is equal to one when the condition inside is satisfied and

zero otherwise. Let J be a random variable distributed according to $\hat{q}$. Let $\mathbb{P}_h$ be the law of J when the problem instance is h and $\mathbb{E}_h$ denote expectations in problem instance h . Note that we have $\mathbb{P}_h(J = V_h) = \mathbb{E}_h \left[\frac{1}{T} \sum_{t=1}^T \mathbf{1}_{\{S_t = V_h\}} \right]$. In addition, for any problem instance $h > 0$, we denote the set of optimal actions by $\mathcal{S}_h^* := \{V_i : \mathbb{E}_h[\sigma(V_i)] = (k-1)\frac{1-\epsilon}{2} + \frac{1+\epsilon}{2}\}$. Hence, for the h -th problem instance, we have

$$R_h(T) := \mathbb{E}_h[R^{(1,1)}(T)] - c\mathbb{E}_h \left[\sum_{t=1}^T B_t \right] \quad (\text{A.28})$$

$$\begin{aligned} &= \mathbb{E}_h \left[\sum_{t=1}^T \mathbf{1}_{\{S_t \notin \mathcal{S}_h^*\}} \epsilon \right] \\ &= \epsilon T \mathbb{E}_h \left[\sum_{t=1}^T \mathbf{1}_{\{S_t \notin \mathcal{S}_h^*\}} / T \right] \\ &= \epsilon T \mathbb{P}_h(J \notin \mathcal{S}_h^*) \end{aligned} \quad (\text{A.29})$$

$$= \epsilon T (1 - \mathbb{P}_h(J \in \mathcal{S}_h^*)). \quad (\text{A.30})$$

Therefore, there exist at least one problem instance for which R_T is greater than or equal to the mean of (A.30) over all problem instances $h > 0$:

$$\sup_{h>0} R_h(T) \geq \epsilon T \left(1 - \frac{1}{m} \sum_{h=1}^m \mathbb{P}_h(J \in \mathcal{S}_h^*) \right). \quad (\text{A.31})$$

Second Step: Pinsker's Inequality

By Pinsker's inequality and union bound, we have

$$\begin{aligned} \mathbb{P}_h(J \in \mathcal{S}_h^*) &\leq \mathbb{P}_0(J \in \mathcal{S}_h^*) + \sqrt{\frac{1}{2} \text{KL}(\mathbb{P}_0 || \mathbb{P}_h)} \\ &\leq \sum_{V_i \in \mathcal{S}_h^*} \mathbb{P}_0(J = V_i) + \sqrt{\frac{1}{2} \text{KL}(\mathbb{P}_0 || \mathbb{P}_h)}. \end{aligned}$$

Finally, by taking the average over all problem instances $h > 0$, and then, applying Jensen's Inequality using concavity of the square root, we obtain

$$(\frac{1}{A} \sum_{i=1}^A \sqrt{x_i} \leq \sqrt{\frac{1}{A} \sum_{i=1}^A x_i} \text{ since } \sqrt{\cdot} \text{ is concave})$$

$$\frac{1}{m} \sum_{h=1}^m \mathbb{P}_h(J \in \mathcal{S}_h^*) \leq \frac{k}{m} + \frac{1}{m} \sum_{h=1}^m \sqrt{\frac{1}{2} \text{KL}(\mathbb{P}_0 || \mathbb{P}_h)} \quad (\text{A.32})$$

$$\leq \frac{k}{m} + \sqrt{\frac{1}{2m} \sum_{h=1}^m \text{KL}(\mathbb{P}_0 || \mathbb{P}_h)} \quad (\text{A.33})$$

where (A.32) is due to the fact that $\sum_h \sum_{V_i \in \mathcal{S}_h^*} \mathbb{P}_0(J = V_i) = k$ since each action V_i is optimal exactly for k problem instances out of m problem instances with $h > 0$.

Third Step: Computation of $KL(\mathbb{P}_0 || \mathbb{P}_h)$

Let Z_t be the set of observed edges and O_t be the set of influence outcomes on the observed edges in epoch t , $O_{1:t} := \{O_1, \dots, O_t\}$, and let $\mathbb{P}_h^t$ be the law of $O_{1:t}$ in problem instance h . Note that the sequence of observations $O_{1:T}$ deterministically determines the sequence of actions $\{S_1, \dots, S_T\}$ taken by the learner, which in turn determines the law of J . Similarly, $O_{1:t-1}$ deterministically determines Z_t . In order to proceed from (A.33), first by the data processing inequality, we obtain $\text{KL}(\mathbb{P}_0 || \mathbb{P}_h) \leq \text{KL}(\mathbb{P}_0^T || \mathbb{P}_h^T)$. Then, we apply the chain rule for KL-divergence to

get

$$\begin{aligned}
\text{KL}(\mathbb{P}_0 || \mathbb{P}_h) &\leq \text{KL}(\mathbb{P}_0^T || \mathbb{P}_h^T) \\
&= \text{KL}(\mathbb{P}_0^1 || \mathbb{P}_h^1) + \sum_{t=2}^T \mathbb{E}_{o_{1:t-1}} [\text{KL}(\mathbb{P}_0^t(\cdot | o_{1:t-1}) || \mathbb{P}_h^t(\cdot | o_{1:t-1}))] \\
&= \text{KL}(\mathbb{P}_0^1 || \mathbb{P}_h^1) \\
&\quad + \sum_{t=2}^T \sum_{o_{1:t-1}} \mathbb{P}_0^{t-1}(o_{1:t-1}) \text{KL}(\mathbb{P}_0^t(\cdot | o_{1:t-1}) || \mathbb{P}_h^t(\cdot | o_{1:t-1})) \\
&= \text{KL}(\mathbb{P}_0^1 || \mathbb{P}_h^1) \\
&\quad + \sum_{t=2}^T \sum_{o_{1:t-1}} \mathbb{P}_0^{t-1}(o_{1:t-1}) \mathbf{1}_{\{(v_0^h, v_1^h) \in Z_t | o_{1:t-1}\}} \text{KL} \left(\frac{1-\epsilon}{2} \middle| \middle| \frac{1+\epsilon}{2} \right) \tag{A.34}
\end{aligned}$$

$$= \text{KL} \left(\frac{1-\epsilon}{2} \middle| \middle| \frac{1+\epsilon}{2} \right) \mathbb{E}_0 \left[\sum_{t=1}^T \mathbf{1}_{\{(v_0^h, v_1^h) \in Z_t\}} \right] \tag{A.35}$$

where (A.34) is due to the fact that KL-divergence is non-zero only when the edge (v_0^h, v_1^h) is observed since the influence probabilities corresponding to the other edges are the same for problem instances 0 and h .

Via the inequality, $\text{KL}(p||q) \leq \frac{(p-q)^2}{q(1-q)}$, thus, we have

$$\begin{aligned}
&\sum_{h=1}^m \text{KL}(\mathbb{P}_0^T || \mathbb{P}_h^T) \\
&= \text{KL} \left(\frac{1-\epsilon}{2} \middle| \middle| \frac{1+\epsilon}{2} \right) \sum_{h=1}^m \mathbb{E}_0 \left[\sum_{t=1}^T \mathbf{1}_{\{(v_0^h, v_1^h) \in Z_t\}} \right] \\
&= \text{KL} \left(\frac{1-\epsilon}{2} \middle| \middle| \frac{1+\epsilon}{2} \right) \mathbb{E}_0 \left[\sum_{t=1}^T \sum_{h=1}^m \mathbf{1}_{\{(v_0^h, v_1^h) \in Z_t\}} \right] \\
&\leq \frac{4\epsilon^2}{1-\epsilon^2} OT \tag{A.36}
\end{aligned}$$

where OT is the expected number of edges observed until epoch T .

Fourth Step: ϵ -tuning

Substituting (A.36) into (A.31) and setting $\epsilon \leq \frac{1}{2}$, we get

$$\max_{h>0} R_h(T) \geq \epsilon T \left(1 - \frac{k}{m} - 4\epsilon \sqrt{\frac{O}{6m}T} \right) \quad (\text{A.37})$$

Letting $\epsilon = \alpha \sqrt{\frac{m}{OT}}$ for some $\alpha > 0$ and $k' = \frac{m}{k}$ we obtain

$$\max_{h>0} R_h(T) \geq \sqrt{\frac{mT}{O}} \left(\alpha - \frac{\alpha}{k'} - \frac{4\alpha^2}{\sqrt{6}} \right). \quad (\text{A.38})$$

For $\alpha = \left(1 - \frac{1}{k'}\right) \frac{\sqrt{6}}{8}$, we have

$$\begin{aligned} \max_{h>0} R_h(T) &\geq \left[\frac{\sqrt{6}}{8} \left(1 - \frac{1}{k'}\right)^2 - \frac{4}{\sqrt{6}} \frac{6}{64} \left(1 - \frac{1}{k'}\right)^2 \right] \sqrt{\frac{mT}{O}} \\ &= \frac{\sqrt{6}}{16} \left(1 - \frac{1}{k'}\right)^2 \sqrt{\frac{mT}{O}}. \end{aligned}$$

Hence, by the definition of $R_h(T)$ in (A.28), we have

$$\mathbb{E}_h[R^{(1,1)}(T)] - cOT \geq \frac{\sqrt{6}}{16} \left(1 - \frac{1}{k'}\right)^2 \sqrt{\frac{mT}{O}} \quad (\text{A.39})$$

for at least one problem instance.

This shows that for at least one problem instance h

$$\begin{aligned} \mathbb{E}_h[R^{(1,1)}(T)] &= R_h(T) + cOT \\ &\geq \frac{\sqrt{6}}{16} k_0 \sqrt{\frac{mT}{O}} + cOT \\ &\geq \max \left\{ 2 \times \left(\frac{\sqrt{6}}{32}\right)^{2/3} k_0^{2/3} c^{1/3} m^{1/3} T^{2/3}, \frac{\sqrt{6}}{16} k_0 \sqrt{T} \right\} \end{aligned} \quad (\text{A.40})$$

where $k_0 = \left(1 - \frac{1}{k'}\right)^2$, by setting $O = \left(\frac{\sqrt{6}}{32}\right)^{2/3} k_0^{2/3} c^{-2/3} m^{1/3} T^{-1/3}$ in (A.40) and

using the fact that O is upper bounded by m .

