Reward Maximization Under Uncertainty: Leveraging Side-Observations on Networks

Swapna Buccapatnam sb646f@att.com
AT&T Labs Research, Middletown, NJ 07748, USA

Fang Liu liu.3977@osu.edu
Atilla Eryilmaz eryilmaz.2@osu.edu
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43210, USA

Ness B. Shroff shroff.11@osu.edu
Department of Electrical and Computer Engineering and Computer Science Engineering
The Ohio State University
Columbus, OH 43210, USA

Editor: Shie Mannor

Abstract

We study the stochastic multi-armed bandit (MAB) problem in the presence of side-observations across actions that occur as a result of an underlying network structure. In our model, a bipartite graph captures the relationship between actions and a common set of unknowns such that choosing an action reveals observations for the unknowns that it is connected to. This models a common scenario in online social networks where users respond to their friends' activity, thus providing side information about each other's preferences. Our contributions are as follows: 1) We derive an asymptotic lower bound (with respect to time) as a function of the bi-partite network structure on the regret of any *uniformly good policy* that achieves the maximum long-term average reward. 2) We propose two policies - a randomized policy; and a policy based on the well-known upper confidence bound (UCB) policies - both of which explore each action at a rate that is a function of its network position. We show, under mild assumptions, that these policies achieve the asymptotic lower bound on the regret up to a multiplicative factor, independent of the network structure. Finally, we use numerical examples on a real-world social network and a routing example network to demonstrate the benefits obtained by our policies over other existing policies.

Keywords: Multi-armed Bandits, Side Observations, Bipartite Graph, Regret Bounds

1. Introduction

Multi-armed bandit (MAB) problems are well-known models of sequential decision-making under uncertainty (Lai and Robbins, 1985) and have lately been used to model new and exciting decision problems in content recommendation systems, online advertising platforms, and social networks, among others. In the classical MAB setting, at each time, a bandit policy must choose an action from a set of actions with unknown probability distributions. Choosing an action gives a random reward drawn from the distribution of the action. The regret of any policy is defined as the difference between the total reward obtained from the action with the highest average reward and the given policy's total reward. The goal is to find policies that minimize the expected regret over time.

In this work, we consider an important extension to the classical MAB problem, where choosing an action not only generates a reward from that action, but also reveals important information for a subset of the remaining actions. We model this relationship between different actions using a bipartite graph between the set of actions and a common set of unknowns (see Figure 2). The reward from each action is a known function of a subset of the unknowns (called its parents) and choosing an action reveals observations from each of its parents. Our main objective in this work is to leverage such a structure to improve scalability of bandit policies in terms of the action/decision space.

Such an information structure between actions becomes available in a variety of applications. For example, consider the problem of *routing* in communication networks, where packets are to be sent over a set of links from source to destination (called a path or a route) in order to minimize the delay. Here, the total delay on each path is the sum of individual link delays, which are unknown. In addition, traveling along a path reveals observations for delays on each of constituent links. Hence, each path provides additional information for all other paths that share some of their links with it. In this example, actions correspond to a set of feasible paths and the set of unknowns corresponds to random delays on all the links in the network.

Another example occurs in advertising in online social networks through promotional offers. Suppose a user is offered a promotion/discounted price for a product in return for advertising it to his friends/neighbors in an online social network. The influence of the user is then measured by the friends that respond to his message through comments/likes, etc. Each user has an intrinsic unknown probability of responding to such messages on social media. Here, the set of actions correspond to the set of users (to whom promotions are given) and the set of unknowns are the users' intrinsic responsiveness to such promotions.

In this work, we aim to characterize the asymptotic lower bound on the regret for a general stochastic multi-armed bandit problem in the presence of such an information structure and investigate policies that achieve this lower bound by taking the network structure into account. Our main contributions are as follows:

- We model the MAB problem in the presence of additional structure and derive an asymptotic (with respect to time) lower bound (as a function of the network structure) on the regret of any uniformly good policy which achieves the maximum long term average reward. This lower bound is presented in terms of the optimal value of a linear program (LP).
- Motivated by the LP lower bound, we propose and investigate the performance of a randomized policy, we call ϵ_t-greedy-LP policy, as well as an upper confidence bound based policy, we call UCB-LP policy. Both of these policies *explore each action at a rate that is a function of its location in the network.* We show under some mild assumptions that these policies are optimal in the sense that they achieve the asymptotic lower bound on the regret up to a multiplicative constant that is independent of the network structure.

The model considered in this work is an important first step in the direction of more general models of interdependence across actions. For this model, we show that as the number of actions becomes large, significant benefits can be obtained from policies that explicitly take network structure into account. While ϵ_t-greedy-LP policy explores actions at a rate proportional to their network position, its exploration is oblivious to the average rewards of the sub-optimal actions. On the other hand, UCB-LP policy takes into account both the upper confidence bounds on the mean rewards as well as network position of different actions at each time.

2. Related Work

The seminal work of Lai and Robbins (1985) showed that the asymptotic lower bound on the regret of any uniformly good policy scales logarithmically with time with a multiplicative constant that is a function of the distributions of actions. Further, Lai and Robbins (1985) provide constructive policies called Upper Confidence Bound (UCB) policies based on the concept of optimism in the face of uncertainty that asymptotically achieve the lower bound. More recently, Auer et al. (2002) considered the case of bounded rewards and propose simpler sample-mean-based UCB policies and a decreasing-ϵ_t-greedy policy that achieve logarithmic regret uniformly over time, rather than only asymptotically as in the previous works.

The traditional multi-armed bandit policies incur a regret that is linear in the number of suboptimal arms. This makes them unsuitable in settings such as content recommendation, advertising, etc, where the action space is typically very large. To overcome this difficulty, richer models specifying additional information across reward distributions of different actions have been studied, such as dependent bandits by Pandey et al. (2007), $\mathcal{X}$-armed bandits by Bubeck et al. (2011), linear bandits by Rusmevichientong and Tsitsiklis (2010), contextual side information in bandit problems by Li et al. (2010), combinatorial bandits by Chen et al. (2013) etc..

The works of Mannor and Shamir (2011), Caron et al. (2012), and Buccapatnam et al. (2014) proposed to handle the large number of actions by assuming that choosing an action reveals observations from a larger set of actions. In this setting, actions are embeded in a network and choosing an action provides observations for all the immediate neighbors in the network. The policies proposed in Mannor and Shamir (2011) achieve the best possible regret in the adversarial setting (see Bubeck and Cesa-Bianchi (2012) for a survey of adversarial MABs) with side-observations, and the regret bounds of these policies are in terms of the independence number of the network. The stochastic version of this problem is introduced in Caron et al. (2012) and Buccapatnam et al. (2014), which improves upon the results in Caron et al. (2012). In Buccapatnam et al. (2014), the authors derive a lower bound on regret in stochastic network setting for any uniformly good policy and propose two policies that achieve this lower bound in these settings up to a multiplicative constant. Our current work extends the setting in Caron et al. (2012); Buccapatnam et al. (2014) to a more general and important graph feedback structure between the set of actions and a set of common unknowns, which may or may not coincide with the set of actions available to the decision maker. The setting of Mannor and Shamir (2011), Caron et al. (2012), and Buccapatnam et al. (2014) is a special case of this general feedback structure, where the set of unknowns and the set of actions coincide.

More recently, Cohen et al. (2016), have studied the multi-armed bandit problem with a graph based feedback structure similar to Mannor and Shamir (2011), and Buccapatnam et al. (2014). However, they assume that the graph structure is never fully revealed. In contrast, in many cases such as the problem of routing in communication networks and the problem of influence maximization in social networks, the graph structure is revealed or learnt apriori and is known. When the graph structure is known, the authors in Buccapatnam et al. (2014) propose algorithms for the stochastic setting whose regret performance is bounded by the domination number of the graph. In contrast, the algorithms proposed in Cohen et al. (2016) assume that the graph is unknown and achieve a regret that is upper bounded by the independence number of the graph. (Note that the independence number of a graph is larger than or equal to the domination number). Our current work proposes a general feedback structure of which Buccapatnam et al. (2014) and Cohen et al. (2016) can be viewed as a special case. Moreover, we present algorithms that benefit significantly from the knowledge of the graph feedback structure.

The setting of combinatorial bandits (CMAB) by Chen et al. (2013) is also closely related to our work. In CMAB, a subset of base actions with unknown distributions form super actions and in each round, choosing a super action reveals outcomes of its constituent actions. The reward obtained is a function of these outcomes. The number of super actions and their composition in terms of base actions is assumed to be arbitrary and the policies do not utilize the underlying network structure between base actions and super actions. In contrast, in our work, we derive a regret lower bound in terms of the underlying network structure and propose policies that achieve this bound. This results in markedly improved performance when the number of super actions is not substantially larger than the number of base actions.

3. Problem Formulation

In this section, we formally define the general bandit problem in the presence of side observations across actions. Let $\mathcal{N} = \{1, \dots, N\}$ denote the collection of *base-arms* with unknown distributions. Subsets of base-arms form *actions*, and are indexed by $\mathcal{K} = \{1, \dots, K\}$. A decision maker must choose an action $j \in \mathcal{K}$ at each time t and observes the rewards of related base-arms. Let $X_i(t)$ be the reward of base-arm i observed by the decision maker (on choosing some action) at time t. We assume that $\{X_i(t), t \geq 0\}$ are independent and identically distributed (i.i.d.) for each i and $\{X_i(t), \forall i \in \mathcal{N}\}$ are independent for each time t. Let $V_j \subseteq \mathcal{N}$ be the subset of base-arms that are observed when playing action j. Then, we define $S_i = \{j : i \in V_j\}$ as the support of base-arm i, i.e., the decision maker gets observations for base-arm i on playing action $j \in S_i$. When the decision maker chooses action j at time t, he or she observes one realization for each of the random variables $X_i(t)$, $i \in V_j$. The reward of the played action j depends on the outcomes of its related base-arms subset, denoted by $\mathcal{K}_j \subseteq \mathcal{N}$, and some known function $f_j(\cdot)$. Note that $\mathcal{K}_j \subseteq V_j$ because there may be some base-arms that can be observed by action j but not counted as reward in general (see Figure 2 for a concrete example). Let the vector $\vec{X}_j(t) = [X_i(t)]_{i \in \mathcal{K}_j}$ denote the collection of random variables associated with the reward of action j. Then the reward from playing action j at time t is given by $f_j(\vec{X}_j(t))$. We assume that the reward is bounded in $[0, 1]$ for each action. Note that we only assume that the reward function $f_j(\cdot)$ is bounded and the specific form of $f_j(\cdot)$ and $\mathcal{K}_j$ are determined by the decision maker or the specific problem. Let μ_j be the mean of reward on playing action j.

Side-observation model : The actions $\mathcal{K}$ and base-arms $\mathcal{N}$ form nodes in a network G, represented by a bipartite graph $(\mathcal{K}, \mathcal{N}, E)$ and the collection $\{\mathcal{K}_j\}_{j\in\mathcal{K}}$. The $N \times K$ adjacency matrix $E = [e_{ij}]$ is defined by $e_{ij} = 1$ if $i \in V_j$ and $e_{ij} = 0$ otherwise. If there is an edge between action j and base-arm i, i.e., $e_{ij} = 1$, then we can observe a realization of base-arm i when choosing action j. Intuitively, the bipartite graph determined by $\{V_j\}_{j\in\mathcal{K}}$ describes the side-observation relationships while the collection $\{\mathcal{K}_j\}_{j\in\mathcal{K}}$ captures the reward structure. Without loss of generality, we assume that $\cup_{i\in\mathcal{K}}\mathcal{K}_i = \mathcal{N}$, which means that there are no useless (dummy) unknown base-arms in the network.

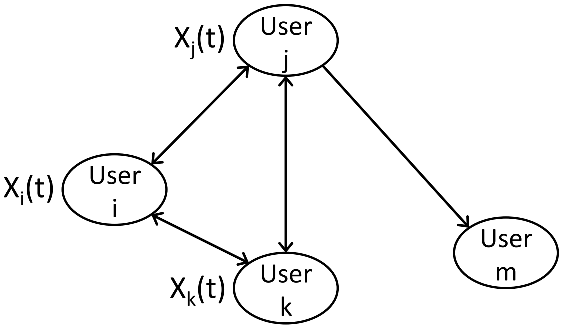

Figure 1: At time t, suppose that the decision maker chooses user i to offer a promotion. He then receives a response $X_i(t)$ from user i. Using the social interconnections, he also observes responses $X_j(t)$ and $X_k(t)$ of i's neighbors j and k.

Figure 1 illustrates the side-observation model for the example of targeting users in an online social network. Such side observations are made possible in settings of online social networks like Facebook by surveying or tracking a user's neighbors' reactions (likes, dislikes, no opinion, etc.) to the user's activity. This is possible when the online social network has a survey or a like/dislike indicator that generates side observations. For example, when user i is offered a promotion, her neighbors may be queried as follows: "User i was recently offered a promotion. Would you also be interested in the offer?[1]"

Figure 2 shows the bipartite graph generated from the example shown in Figure 1. The set of base-arms is the set of users since they act independently according to their own preferences in the promotion, which are unknown to the decision maker. The set of actions is also the set of users because the decision maker wants to target the users with the maximum expected reward. When action j (user j) is chosen, the decision maker observes $X_i(t)$, $X_j(t)$, $X_k(t)$ and $X_m(t)$ from user i, j, k and m since $V_j = \{i, j, k, m\}$. The reward of playing action j depends on $\mathcal{K}_j$ and $f_j(\cdot)$. Suppose $f_j(\vec{X}_j(t)) = \sum_{i\in\mathcal{K}_j} X_i(t)$. Then $\mathcal{K}_j = \{i, j, k, m\}$ means that the reward is the sum of all positive feedbacks. It is also

[1] Since, the neighbors do not have any information on whether the user i accepted the promotion, they act independently according to their own preferences in answering this survey. The network itself provides a better way for surveying and obtaining side observations.

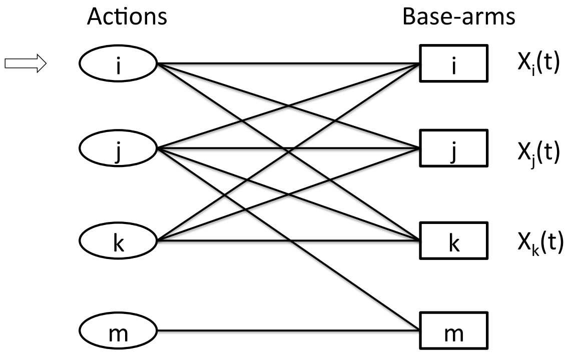

Figure 2: Bipartite graph for the example of targeting users in online social network.

possible that the decision maker set $\mathcal{K}_j = \{j\}$, which means that the reward of playing action j is only the observation from the user j.

The reward function can be quite general (but bounded) to accommodate different settings. Also, the bipartite graph can be more general than social networks in two key ways: 1) Connecting two users in two-hop neighborhood means that the reaction of the friend of my friend is also observable, which is true in Facebook. 2) Connecting two users, say i and j, with similar preference profiles means that the network actively recommends the promotion received by user i to user j even though they are not friends. This has been widely applied in recommender systems such as Yelp.

Objective: An allocation strategy or policy $\boldsymbol{\phi}$ chooses the action to be played at each time. Formally, $\boldsymbol{\phi}$ is a sequence of random variables $\{\phi(t), t \geq 0\}$, where $\phi(t) \in \mathcal{K}$ is the action chosen by policy $\boldsymbol{\phi}$ at time t. Let $T_j^{\phi}(t)$ be the total number of times action j is chosen up to time t by policy $\boldsymbol{\phi}$. For each action, rewards are only obtained when the action is chosen by the policy (side-observations do not contribute to the total reward). Then, the regret of policy $\boldsymbol{\phi}$ at time t for a fixed $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ is defined by

$$R_{\boldsymbol{\mu}}^{\phi}(t) = \mu^* t - \sum_{j=1}^{K} \mu_j \mathbb{E}[T_j^{\phi}(t)] = \sum_{j=1}^{K} \Delta_j \mathbb{E}[T_j^{\phi}(t)],$$

where $\Delta_j \triangleq \mu^* - \mu_j$ and $\mu^* \triangleq \max_{j \in \mathcal{K}} \mu_j$. Henceforth, we drop the superscript ϕ unless it is required. The objective is to find policies that minimize the rate at which the regret grows as a function of time for every fixed network G. We focus our investigation on the class of uniformly good policies (Lai and Robbins, 1985) defined below:

Uniformly good policies: An allocation rule $\boldsymbol{\phi}$ is said to be uniformly good if for every fixed $\boldsymbol{\mu}$, the following condition is satisfied as $t \to \infty$:

$$R_{\boldsymbol{\mu}}(t) = o(t^b), \text{ for every } b > 0.$$

The above condition implies that uniformly good policies achieve the optimal long term average reward of μ^*. Next, we define two structures that will be useful later to bound the performance of allocation strategies in terms of the network structure G.

Definition 1 *A hitting set D is a subset of $\mathcal{K}$ such that $S_i \cap D \neq \emptyset$, $\forall i \in \mathcal{N}$. Then the hitting set number is $\gamma(G) = \inf_{D \subseteq \mathcal{K}}\{|D| : S_i \cap D \neq \emptyset, \ \forall i \in \mathcal{N}\}$. For example, the set $\{i, m\}$ is a hitting set in Figure 2.*

Definition 2 *A clique C is a subset of $\mathcal{K}$ such that $\mathcal{K}_j \subseteq V_i$, $\forall i, j \in C$. This means that for every action i in C, we can observe the reward of playing any action j in C. A clique cover $\mathcal{C}$ of a network G is a partition of all its nodes into sets $C \in \mathcal{C}$ such that the sub-network formed by each C is a clique. Let $\bar{\chi}(G)$ be the smallest number of cliques into which the nodes of the network G can be partitioned, also called the clique partition number.*

Proposition 3 *For any network G with bipartite graph $(\mathcal{K}, \mathcal{N}, E)$ and $\{\mathcal{K}_j\}_{j \in \mathcal{K}}$, if $\cup_{j \in \mathcal{K}} \mathcal{K}_j = \mathcal{N}$, then $\gamma(G) \leq \bar{\chi}(G)$.*

Proof Let $\mathcal{C} = \{C_1, C_2, ..., C_m\}$ be a clique cover with cardinality m, i.e., $|\mathcal{C}| = m$ and each C_k is a clique for $k = 1, ..., m$. Pick arbitrarily an element a_k from C_k for each k. Define $\mathcal{H} = \{a_k : k = 1, ..., m\}$. Now it remains to show that $\mathcal{H}$ is a hitting set, which implies $\gamma(G) \leq \bar{\chi}(G)$. We prove this by contradiction.

Suppose $\mathcal{H}$ is not a hitting set, then $\exists i \in \mathcal{N}$ s.t. $S_i \cap \mathcal{H} = \emptyset$. Since $\cup_{j \in \mathcal{K}} \mathcal{K}_j = \mathcal{N}$, $\exists j \in \mathcal{K}$ s.t. $i \in \mathcal{K}_j$. $\mathcal{C}$ is a clique cover, then $\exists k(j) \in \{1, 2, ..., m\}$ such that $j \in C_{k(j)}$. By the construction of $\mathcal{H}$, there exists $a_{k(j)} \in \mathcal{H} \cap C_{k(j)}$. By the definition of clique, we have $\mathcal{K}_j \subseteq V_{a_{k(j)}}$. Thus, we have $a_{k(j)} \in S_i$ since $i \in \mathcal{K}_j$. It follows that $S_i \cap \mathcal{H} \neq \emptyset$, which contradicts to $S_i \cap \mathcal{H} = \emptyset$. Hence, $\mathcal{H}$ is a hitting set. ■

In the next section, we obtain an asymptotic lower bound on the regret of uniformly good policies for the setting of MABs with side-observations. This lower bound is expressed as the optimal value of a linear program (LP), where the constraints of the LP capture the connectivity of each action in the network.

4. Regret Lower Bound in the Presence of Side Observations

In order to derive a lower bound on the regret, we need some mild regularity assumptions (Assumptions 1, 2, and 3) on the distributions F_i (associated with base-arm i) that are similar to the ones in Lai and Robbins (1985). Let the probability distribution F_i have a univariate density function $g(x; \theta_i)$ with unknown parameters θ_i, for each $i \in \mathcal{N}$. Let $D(\theta || \sigma)$ denote the Kullback Leibler (KL) distance between distributions with density functions $g(x; \theta)$ and $g(x; \sigma)$ and with means $u(\theta)$ and $u(\sigma)$ respectively.

Assumption 1 *(Finiteness) We assume that $g(\cdot; \cdot)$ is such that $0 < D(\theta || \sigma) < \infty$ whenever $u(\sigma) > u(\theta)$.*

Assumption 2 *(Continuity) For any $\epsilon > 0$ and θ, σ such that $u(\sigma) > u(\theta)$, there exists $\eta > 0$ for which $|D(\theta || \sigma) - D(\theta || \rho)| < \epsilon$ whenever $u(\sigma) < u(\rho) < u(\sigma) + \eta$.*

Assumption 3 *(Denseness) For each $i \in \mathcal{N}$, $\theta_i \in \Theta$ where the set Θ satisfies: for all $\theta \in \Theta$ and for all $\eta > 0$, there exists $\theta' \in \Theta$ such that $u(\theta) < u(\theta') < u(\theta) + \eta$.*

Let $\vec{\theta}$ be the vector $[\theta_1, \dots, \theta_N]$. Define $\Theta_i = \{\vec{\theta} : \exists k \in S_i \text{ such that } \mu_k(\vec{\theta}) < \mu^*(\vec{\theta})\}$. So, not all actions that support base-arm i are optimal. Suppose $\vec{\theta} \in \Theta_i$. For base arm i, define the set

$$\mathcal{B}_i(\theta_i) = \{\theta_i' : \exists k \in S_i \text{ such that } \mu_k(\vec{\theta}_i') > \mu^*(\vec{\theta})\},$$

where $\vec{\theta}_i' = [\theta_1, \dots, \theta_i', \dots \theta_N]$. $\vec{\theta}_i'$ differs from $\vec{\theta}$ only in the i^{th} parameter. In this set $\mathcal{B}_i(\theta_i)$, base-arm i contributes towards a unique optimal action. Define constant $J_i(\theta_i) = \inf\{D(\theta_i||\theta_i') : \theta_i' \in \mathcal{B}_i(\theta_i)\}$. This is well-defined when $\mathcal{B}_i(\theta_i) \neq \emptyset$.

The following proposition is obtained using Theorem 2 in Lai and Robbins (1985). It provides an asymptotic lower bound on the regret of any uniformly good policy under the model described in Section 3:

Proposition 4 *Suppose Assumptions 1, 2, and 3 hold. Let $\mathcal{U} = \{j : \mu_j < \mu^*\}$ be the set of suboptimal actions. Also, let $\Delta_j = \mu^* - \mu_j$. Then, under any uniformly good policy $\boldsymbol{\phi}$, the expected regret is asymptotically bounded below as follows:*

$$\liminf_{t \to \infty} \frac{R_{\boldsymbol{\mu}}(t)}{\log(t)} \geq c_{\boldsymbol{\mu}}, \tag{1}$$

where $c_{\boldsymbol{\mu}}$ is the optimal value of the following linear program (LP) P_1:

$$\begin{aligned} P_1 : \ \min \sum_{j \in \mathcal{U}} \Delta_j w_j, \\ \textit{subject to: } \sum_{j \in S_i} w_j \geq \frac{1}{J_i(\theta_i)}, \ \forall i \in \mathcal{N}, \\ w_j \geq 0, \ \forall j \in \mathcal{K}. \end{aligned}$$

Proof *(Sketch)* Let $M_i(t)$ be the total number of observations corresponding to base-arm i available at time t. Then, by modifying the proof of Theorem 2 of Lai and Robbins (1985), we have that, for $i \in \mathcal{N}$,

$$\liminf_{t \to \infty} \frac{\mathbb{E}[M_i(t)]}{\log(t)} \geq \frac{1}{J_i(\theta_i)}.$$

An observation is received for base-arm i whenever any action in S_i is chosen. Hence, $M_i(t) = \sum_{j \in S_i} T_j(t)$. These two facts give us the constraints in LP P_1. See Appendix A for the full proof. ■

The linear program given in P_1 contains the graphical information that governs the lower bound. However, it requires the knowledge of $\vec{\theta}$, which is unknown. This motivates the construction of the following linear program, LP P_2, which preserves the graphical structure while eliminating the distributional dependence on $\vec{\theta}$.

$$\begin{aligned} P_2 : \min \sum_{j \in \mathcal{K}} z_j \\ \text{subject to: } \sum_{j \in S_i} z_j \geq 1, \ \forall i \in \mathcal{N}, \\ \text{and } z_j \geq 0, \ \forall j \in \mathcal{K}. \end{aligned}$$

Let $\mathbf{z}^* = (z_j^*)_{j \in \mathcal{K}}$ be the optimal solution of LP P_2. In Sections 5 and 6, we use the above LP P_2 to modify the ϵ-greedy policy in Auer et al. (2002) and UCB policy in Auer and Ortner (2010) for the setting of side-observations. We provide regret guarantees of these modified policies in terms of the optimal value $\sum_{j \in \mathcal{K}} z_j^*$ of LP P_2. We note that the linear program P_2 is, in fact, the LP relaxation of the minimum hitting set problem on network G. Since, any hitting set of network G is a feasible solution to the LP P_2, we have that the optimal value of the LP $\sum_{j \in \mathcal{K}} z_j^* \leq \gamma(G) \leq \bar{\chi}(G)$.

Proposition 5 *Consider an Erdos-Renyi random bipartite graph $(\mathcal{K}, \mathcal{N}, E)$ such that each entry of the matrix E equals 1 with probability p, where $0 < p < 1$. Suppose $\cup_{j \in \mathcal{K}} \mathcal{K}_j = \mathcal{N}$, i.e., there are no useless base-arms in the network, then $\sum_{j \in \mathcal{K}} z_j^*$ is upper-bounded by $\log_{\frac{1}{1-p}} N$ as $N \to \infty$ in probability.*

Proof *(sketch)* Since $\sum_{j \in \mathcal{K}} z_j^* \leq \gamma(G)$, it remains to be shown that $\gamma(G)$ is upper bounded by the above result. Suppose there are no useless base-arms in the network. Then the set of all actions is a hitting set. Based on this observation, we construct a repeated experiment to generate actions sequentially. Then we define a stopping time τ as the first time that all the generated actions form a hitting set. Hence, we show the asymptotic result of τ as the upper bound of $\gamma(G)$. See full proof in Appendix B. ■

In the next proposition, we provide a lower bound on $c_{\boldsymbol{\mu}}$ in Equation (1) using the optimal solution $\mathbf{z}^* = (z_j^*)_{j \in \mathcal{K}}$ of LP P_2.

Proposition 6 *Let $\mathcal{U} = \{j : \mu_j < \mu^*\}$ be the set of suboptimal actions. Let $\mathcal{O} = \{j : \mu_j = \mu^*\}$ be the set of optimal actions. Then,*

$$\frac{\max_{i \in \mathcal{N}} J_i(\theta_i)}{\min_{j \in \mathcal{U}} \Delta_j} c_{\boldsymbol{\mu}} + |\mathcal{O}| \geq \sum_{j \in \mathcal{K}} z_j^* \geq \frac{\min_{i \in \mathcal{N}} J_i(\theta_i)}{\max_{j \in \mathcal{U}} \Delta_j} c_{\boldsymbol{\mu}}. \tag{2}$$

Proof *(Sketch)* Using the optimal solution of LP P_1, we construct a feasible solution satisfying constraints in LP P_2 for base-arms in $\mathcal{N}$. The feasible solution constructed in this way gives an upper bound on the optimal value of LP P_2 in terms of the optimal value of LP P_1. For the lower bound, we use the fact that any feasible solution of P_2, in particular $\mathbf{z}^*$, can be used to construct a feasible solution of P_1. See Appendix C for the full proof. ■

We note that $\sum_{j \in \mathcal{K}} z_j^* = \Theta(c_\mu)$ completely captures the time dependence of the regret on network structure under the following assumption:

Assumption 4 *The quantities $|\mathcal{O}|$, $\min_{j \in \mathcal{U}} \Delta_j$, and $\min_{i \in \mathcal{N}} J_i(\theta_i)$ are constants that are independent of network size K and N.*

Note that the constants in the above assumption are unknown to the decision maker. In the next section, we propose the ϵ_t-greedy-LP policy which achieves the regret lower bound of $c_{\boldsymbol{\mu}} \log(t)$ up to a multiplicative constant factor that is independent of the network structure and time.

5. Epsilon-greedy-LP policy

Motivated by the LPs P_1 and P_2, we propose a *network-aware* randomized policy called the ϵ_t-greedy-LP policy. We provide an upper bound on the regret of this policy and show that it achieves the asymptotic lower bound, up to a constant multiplier, independent of the network structure. Let $\bar{f}_j(t)$ be the empirical average of observations (rewards and side-observations combined) available for action j up to time t. The ϵ_t-greedy-LP policy is described in Algorithm 1. The policy consists of two iterations - exploitation and exploration, where the exploration probability decreases as $1/t$, similarly to that of the ϵ_t-greedy policy proposed by Auer et al. (2002). However, in our policy, we choose the exploration probability for action j to be proportional to z_j^*/t, where $\mathbf{z}^*$ is the optimal solution of LP P_2, while in the original policy in Auer et al. (2002), the exploration probability is uniform over all actions.

Algorithm 1 : ϵ_t-greedy-LP

```
Input: $c > 0$, $0 < d < 1$, optimal solution $\mathbf{z}^*$ of LP $P_2$.
  for each time $t$ do
    Update $\bar{f}_j(t)$ for each $j \in \mathcal{K}$, where $\bar{f}_j(t)$ is the empirically average over all the observations of action $j$.
    Let $\epsilon(t) = \min\left(1, \frac{c\sum_{j\in\mathcal{K}} z_j^*}{d^2 t}\right)$ and $a^* = \arg\max_{j\in\mathcal{K}} \bar{f}_j(t)$.
    Sample $a$ from the distribution such that $\mathbb{P}\{a = j\} = \frac{z_j^*}{\sum_{i\in\mathcal{K}} z_i^*}$ for all $j \in \mathcal{K}$.
    Play action $\phi(t)$ such that
        $\phi(t) = \begin{cases} a, & \text{with probability } \epsilon(t) \\ a^*, & \text{with probability } 1-\epsilon(t) \end{cases}$    (3)
  end for
```

The following proposition provides performance guarantees on the ϵ_t-greedy-LP policy:

Proposition 7 *For $0 < d < \min_{j\in\mathcal{U}} \Delta_j$, any $c > 0$, and $\alpha > 1$, the probability with which a suboptimal action j is selected by the ϵ_t-greedy-LP policy, described in Algorithm 1, for all $t > t' = \frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2}$ is at most*

$$\left(\frac{c}{d^2 t} z_j^*\right) + \frac{2\lambda c\delta}{\alpha d^2}\left(\frac{et'}{t}\right)^{cr/\alpha d^2} \log\left(\frac{e^2 t}{t'}\right) + \frac{4}{\Delta_j^2}\left(\frac{et'}{t}\right)^{\frac{c\Delta_j^2}{2\alpha d^2}}, \tag{4}$$

where $r = \frac{3(\alpha-1)^2}{8\alpha-2}$, $\lambda = \max_{j\in\mathcal{K}} |\mathcal{K}_j|$, and $\delta = \max_{i\in\mathcal{N}} |S_i|$ is the maximum degree of the supports in the network. Note that α is a parameter we introduce in the proof, which is used to determine a range for the choice of parameter c as shown in Corollary 8.

Proof *(Sketch)* Since $\mathbf{z}^*$ satisfies the constraints in LP P_2, there is sufficient exploration within each suboptimal action's neighborhood. The proof is then a combination of this fact and the proof of Theorem 3 in Auer et al. (2002). In particular, we derive an upper bound for the probability that suboptimal action j is played at each time and then sum over the time. See Appendix D for the full proof. ■

In the above proposition, for large enough c, we see that the second and third terms are $O(1/t^{1+\epsilon})$ for some $\epsilon > 0$ (Auer et al., 2002). Using this fact, the following corollary bounds the expected regret of the ϵ_t-greedy-LP policy:

Corollary 8 *Choose parameters c and d such that,*

$$0 < d < \min_{j \in \mathcal{U}} \Delta_j, \quad \textit{and} \quad c > \max(2\alpha d^2/r, 4\alpha),$$

for any $\alpha > 1$. Then, the expected regret at time T of the ϵ_t-greedy-LP policy described in Algorithm 1 is at most

$$\left(\frac{c}{d^2} \sum_{j \in \mathcal{U}} \Delta_j z_j^* \right) \log(T) + O(K), \tag{5}$$

where the $O(K)$ term captures constants independent of time but dependent on the network structure. In particular, the $O(K)$ term is at most

$$\sum_{j \in \mathcal{U}} \left[\frac{\pi^2 \lambda c \delta \Delta_j}{3 \alpha d^2} \left(e t' \right)^{cr/\alpha d^2} + \frac{2\pi^2}{3\Delta_j} \left(e t' \right)^{\frac{c \Delta_j^2}{2 \alpha d^2}} \right],$$

where t', r, λ and δ are defined in Proposition 7.

Remark 9 *Under Assumption 4, we can see from Proposition 6 and Corollary 8 that, ϵ_t-greedy-LP algorithm is* order optimal *achieving the lower bound $\Omega \left(\sum_{j \in \mathcal{K}} z_j^* \log(T) \right) = \Omega \left(c_{\boldsymbol{\mu}} \log(T) \right)$ as the network and time scale.*

While the ϵ_t-greedy-LP policy is network aware, its exploration is oblivious to the observed average rewards of the sub-optimal actions. Further, its performance guarantees depend on the knowledge of $\min_{j \in \mathcal{U}} \Delta_j$, which is the difference between the best and the second best optimal actions. On the other hand, the UCB-LP policy proposed in the next section is network-aware taking into account the average rewards of suboptimal actions. This could lead to better performance compared to ϵ_t-greedy-LP policy in certain situations, for example, when the action with greater z_j^* is also highly suboptimal.

6. UCB-LP policy

In this section we develop the UCB-LP policy defined in Algorithm 2 and obtain upper bounds on its regret. The UCB-LP policy is based on the improved UCB policy proposed in Auer and Ortner (2010), which can be summarized as follows: the policy estimates the values of Δ_i in each round by a value $\tilde{\Delta}_m$ which is initialized to 1 and halved in each round m. By each round m, the policy draws $n(m)$ observations for each action in the set of actions not eliminated by round m, where $n(m)$ is determined by $\tilde{\Delta}_m$. Then, it eliminates those actions whose UCB indices perform poorly. Our policy differs from the one in Auer and Ortner (2010) by accounting for the presence of side-observations - this is achieved by choosing each action according to the optimal solution of LP P_2, while ensuring that $n(m)$ observations are available for each action not eliminated by round m.

```
**Algorithm 2** : UCB-LP policy
**Input**: Set of actions $\mathcal{K}$, time horizon $T$, and optimal solution $\mathbf{z}^*$ of LP $P_2$.
**Initialization**: Let $\tilde{\Delta}_0 := 1$, $A_0 := \mathcal{K}$, and $B_0 := \mathcal{K}$
  **for** round $m = 0, 1, 2, \dots, \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$ **do**
    **Action Selection:** Let $n(m) := \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$
    **If** $|B_m| = 1$: choose the single action in $B_m$ until time $T$.
    **Else If** $\sum_{i \in \mathcal{K}} z_i^* \leq 2|B_m|\tilde{\Delta}_m : \forall j \in A_m$, choose action $j$ $\left\lceil z_j^*(n(m) - n(m-1)) \right\rceil$ times.
    **Else** For each action $j$ in $B_m$, choose $j$ for $[n(m) - n(m-1)]$ times.
    Update $\bar{f}_j(m)$ and $T_j(m)$ for each $j \in \mathcal{K}$, where $\bar{f}_j(m)$ is the empirical average reward of action $j$, and $T_j(m)$ is the total number of observations for action $j$ up to round $m$.
    **Action Elimination:**
    To get $B_{m+1}$, delete all actions $j$ in $B_m$ for which
      $$\bar{f}_j(m) + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2T_j(m)}} < \max_{a \in B_m} \left\{ \bar{f}_a(m) - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2T_a(m)}} \right\},$$
    **Reset:**
    The set $A_{m+1}$ is given as $A_{m+1} = \bigcup_{i \in D_{m+1}} S_i$, where $D_{m+1} = \bigcup_{j \in B_{m+1}} \mathcal{K}_j$.
    Let $\tilde{\Delta}_{m+1} = \frac{\tilde{\Delta}_m}{2}$.
  **end for**
```

The following proposition provides performance guarantees on the expected regret due to UCB-LP policy:

Proposition 10 *For action j, define round m_j as follows:*

$$m_j := \min \left\{ m : \tilde{\Delta}_m < \frac{\Delta_j}{2} \right\}.$$

Define $\bar{m} = \min\left\{ m : \sum_{j\in\mathcal{K}} z_j^* > \sum_{j:m_j>m} 2^{-m+1} \right\}$ *and the set* $B = \{j \in \mathcal{U} : m_j > \bar{m}\}$*. Then, the expected regret due to the UCB-LP policy described in Algorithm 2 is at most*

$$\sum_{j\in\mathcal{U}\setminus B} \Delta_j z_j^* \frac{32\log(T\hat{\Delta}_j^2)}{\hat{\Delta}_j^2} + \sum_{j\in B} \frac{32\log(T\Delta_j^2)}{\Delta_j} + O(K^2), \tag{6}$$

where $\hat{\Delta}_j = \max\{2^{-\bar{m}+2}, \min_{a:j\in G_a}\{\Delta_a\}\}$*,* $G_a = \cup_{i\in\mathcal{K}_a} S_i$*, and* (z_j^*) *is the solution of LP* P_2*. The* $O(K^2)$ *term captures constants independent of time. Further, under Assumption 4, the regret is also at most*

$$O\left(\sum_{j\in\mathcal{K}} z_j^* \log(T)\right) + O(K^2), \tag{7}$$

where (z_j^*) *entirely captures the time dependence on network structure.*

Proof *(Sketch)* The $\log(T)$ term in the regret follows from the fact that, with high probability, each suboptimal action j is eliminated (from the set B_m) on or before the first round m such that $\tilde{\Delta}_m < \Delta_j/2$. See Appendix E for the full proof. ■

Remark 11 *While* ϵ_t*-greedy-LP does not require knowledge of the time horizon* T*, UCB-LP policy requires the knowledge of T. UCB-LP policy can be extended to the case of an unknown time horizon similar to the suggestion in Auer and Ortner (2010). Start with* $T_0 = 2$ *and at end of each* T_l*, set* $T_{l+1} = T_l^2$*. The regret bound for this case is shown in Proposition 19 in Appendix F.*

Next, we briefly describe the policies UCB-N and UCB-MaxN proposed in Caron et al. (2012). In the UCB-N policy, at each time, the action with the highest UCB index is chosen similar to UCB1 policy in Auer et al. (2002). In UCB-MaxN policy, at each time t, the action i with the highest UCB index is identified and its neighboring action j with the highest empirical average reward at time t is chosen.

Remark 12 *The regret upper bound of UCB-N policy is*

$$\inf_{\mathcal{C}} \sum_{C\in\mathcal{C}} \frac{8\max_{j\in C}\Delta_j}{\min_{j\in C}\Delta_j^2}\log(T) + O(K),$$

where $\mathcal{C}$ *is a clique covering of the sub-network of suboptimal actions. The regret upper bound for UCB-MaxN is the same as that for UCB-N with an* $O(|\mathcal{C}|)$ *term instead of the time-invariant* $O(K)$ *term. We show a better regret performance for UCB-LP policy and* ϵ_t*-greedy-LP policies with respect to the* $\log(T)$ *term because* $\sum_{i\in\mathcal{K}} z_i^* \le \gamma(G) \le \bar{\chi}(G)$*. However, the time-invariant term in our policies is* $O(K)$ *and* $O(K^2)$*, can be worse than the time-invariant term* $O(|\mathcal{C}|)$ *in UCB-MaxN.*

Remark 13 *All uniformly good policies that ignore side-observations incur a regret that is at least $\Omega(|\mathcal{U}|\log(t))$ Lai and Robbins (1985), where $|\mathcal{U}|$ is the number of suboptimal actions. This could be significantly higher than the guarantees on the regret of both ϵ_t-greedy-LP policy and UCB-LP policy for a rich network structure as discussed in Remark 12.*

Remark 14 *In our model, we assumed that the side observations are always available. However, in reality, side observations may only be obtained sporadically. Suppose that when action j is chosen, side-observations of base-arms $i \in \mathcal{K}_j$ are obtained almost surely and that of base-arms $i \in V_j \setminus \mathcal{K}_j$ are obtained with a known probability p_j. In this case, Proposition 4 holds with the replacement of LP P_1 with LP P_1' as follows:*

$$\begin{aligned} P_1':\ & \min \sum_{j\in\mathcal{U}} \Delta_j w_j, \\ \textit{subject to: } & \sum_{j\in S_i} (w_j \mathbb{1}_{\{i\in\mathcal{K}_j\}} + p_j w_j \mathbb{1}_{\{i\notin\mathcal{K}_j\}}) \geq \frac{1}{J_i(\theta_i)},\ \forall i \in \mathcal{N}, \\ & w_j \geq 0,\ \forall j \in \mathcal{K}. \end{aligned}$$

Both of our policies work for this setting by changing the LP P_2 to P_2' as follows:

$$\begin{aligned} P_2':\ & \min \sum_{j\in\mathcal{K}} z_j \\ \textit{subject to: } & \sum_{j\in S_i} (z_j \mathbb{1}_{\{i\in\mathcal{K}_j\}} + p_j z_j \mathbb{1}_{\{i\notin\mathcal{K}_j\}}) \geq 1,\ \forall i \in \mathcal{N}, \\ \textit{and } & z_j \geq 0,\ \forall j \in \mathcal{K}. \end{aligned}$$

The regret bounds of our policies will now depend on the optimal solution of LP P_2'.

7. Numerical Results

7.1 Algorithm Performance on Data Trace

We consider the Flixster network dataset for the numerical evaluation of our algorithms. The authors in Jamali and Ester (2010) collected this social network data, which contains about 1 million users and 14 million links. We use graph clustering by Dhillon et al. (2007) to identify two strongly clustered sub-networks of sizes 1000 and 2000 nodes. Both these sub-networks have a degree distribution that is a straight line on a log-log plot indicating a power law distribution commonly observed in social networks. [2]

Our empirical setup is as follows. Let $\mathcal{N}$ be the set of users and $\mathcal{K} = \mathcal{N}$. To be specific, each user in the network is offered a promotion at each time, and accepts the promotion with probability $\mu_i \in [0.3, 0.9]$. Let S_i be the set of one-hop neighbors in the social network of user i (including user i). This is the setting when the Flixster has a

[2] We note that the social network of interest may or may not display a power law behavior. We find that the subgraphs of the Flixster network have a degree distribution that is a straight line on a log-log plot indicating a power law distribution display while the authors in Ugander et al. (2011) show that the degree distribution of the global Facebook network is not a straight line on log-log plot.

survey or a like/dislike indicator that generates side observations of user's neighborhood. Let $\mathcal{K}_j = \{j\}$ and $f_j(X_j) = X_j$, which means that the decision maker receives a random reward of 1 if the chosen user j accepts the promotion or 0 reward otherwise. μ_j is chosen uniformly at random from $[0.3, 0.8]$ and there are 50 randomly chosen users with optimal $\mu_j = 0.9$.

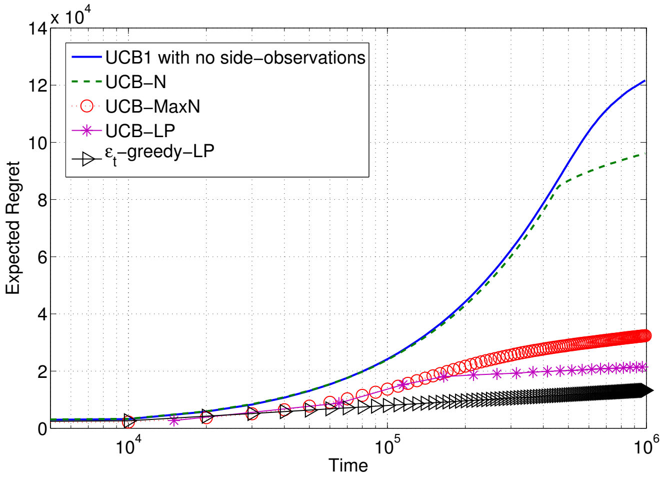

Figure 3: Regret comparison of all the policies for a network of size 1000.

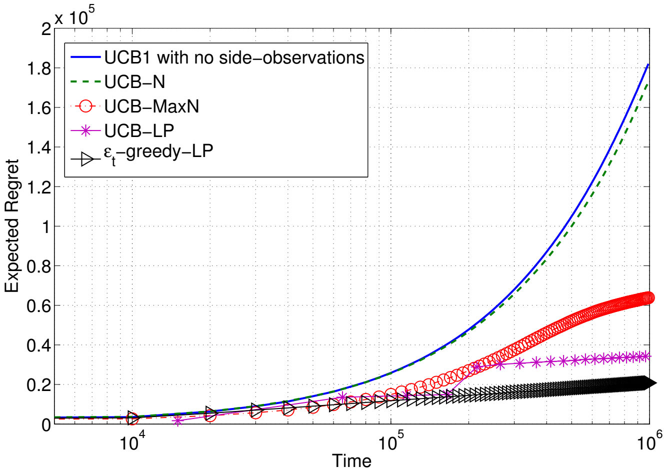

Figure 4: Regret comparison of all the policies for a network of size 2000.

Figures 3 and 4 show the regret performance as a function of time for the two subnetworks of sizes 1000 and 2000 respectively. Note that the average regret is taken over 1000 experiments. For the ϵ_t-greedy-LP policy, we let $c = 5$ and $d = 0.2$ (we choose $d = 0.2$ to show that our algorithm seems to have good performance in more general settings, even when the bounds in the Proposition 7 are not known or used). For both networks, we see that our policies outperform the UCB-N and UCB-MaxN policies Caron et al. (2012)

(UCB-N and UCB-MaxN policies can also be viewed as special cases of those proposed in Chen et al. (2013) for this specific combinatorial structure). We also observe that the improvement obtained by UCB-N policy over the baseline UCB1 policy is marginal. It has been shown (Cooper et al., 2005) that for power law graphs, both $\gamma(G)$ and $\bar{\chi}(G)$ scale linearly with N, although $\gamma(G)$ has a lower slope. Our numerical results show that our policies outperform existing policies even for the Flixster network.

As we show in Corollary 8 and Proposition 10, ϵ_t-greedy-LP and UCB-LP have the same upper bound $O(\sum_{j\in\mathcal{U}} z_j^* \log T)$. It is hard to say which algorithm outperforms the other one. In the Flixster network, we see that the ϵ_t-greedy-LP policy performs better than the UCB-LP policy. As we show in Section 7.2, UCB-LP performs better than ϵ_t-greedy-LP. In addition, the regret gap between the ϵ_t-greedy-LP and UCB-LP is not large compared to their gain to UCB-N and UCB-maxN.

7.2 Algorithm Performance on Synthetic Data

We consider a routing problem defined on a communication network, which is demonstrated as an undirected graph consisting of 6 nodes in Figure 5. We assume that node 1 is the source node and node 6 is the destination node. The decision maker repeatedly sends packets from the source node to the destination node. There exist 13 simple paths from the source node to the destination node. The delay of each path is the sum of delays over all the constituent links. The goal of the decision maker is to identify the path with the smallest expected delay and minimize the regret as much as possible.

Solving the routing problem is a direct application of this work once we let the set of paths be the set of actions and the set of links be the set of base-arms. We assume that the delay of each link i, denoted by $X_i(t)$, is an independent and identically distributed sequence of random variables (drawn from the steady-state distribution) over discrete time t. Then, this is a stochastic bandit problem with side-observations since playing one action (path) reveals some observations of some base-arms (links) that contribute to other actions (paths). For example, choosing path $(1, 2, 4, 6)$ reveals the delay performance of link $(1, 2)$ and $(2, 4)$, which are included in the path $(1, 2, 4, 5, 6)$.

In the routing problem, there are 13 paths and 12 directed links (note that some links are never used in all the paths). Thus, we set $K = 13$ and $N = 12$ in the simulation. Then, we construct the set V_j for each action j such that $i \in V_j$ if path j traverses the link i. And, we set $\mathcal{K}_j = V_j$ for each $j \in \mathcal{K}$ since the delay of a path is the total delay of the traversed links. Let B be the upper bound of all the action delays. Then, we choose the function $f_j(\vec{X}_j(t)) = 1 - \sum_{i\in\mathcal{K}_j} X_i(t)/B$ as the reward of playing action j at time t. In the simulation, we assume that the delay of link i is sampled from a Bernoulli distribution with mean u_i. Each u_i is independently sampled from a uniform distribution from 0 to 1, which realizes the problem instance in Figure 5[3]. One can check that the optimal action (shortest path) is the path $(1, 3, 5, 6)$ given the ground truth $\{u_i\}_{i\in\mathcal{N}}$. We let $B = 5$ in the simulation.

We apply the UCB1, UCB-N, Cohen (Cohen et al., 2016) and our policies to the problem instance in Figure 5 and the regret performance, averaged over 1000 experiments, is shown in Figure 6. We do not represent the result of the UCB-MaxN because it degenerates to

[3] The number indicates the mean delay and the arrow indicates the direction of the link

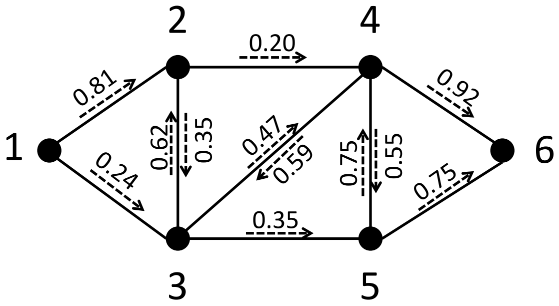

Figure 5: Routing problem on a communication network

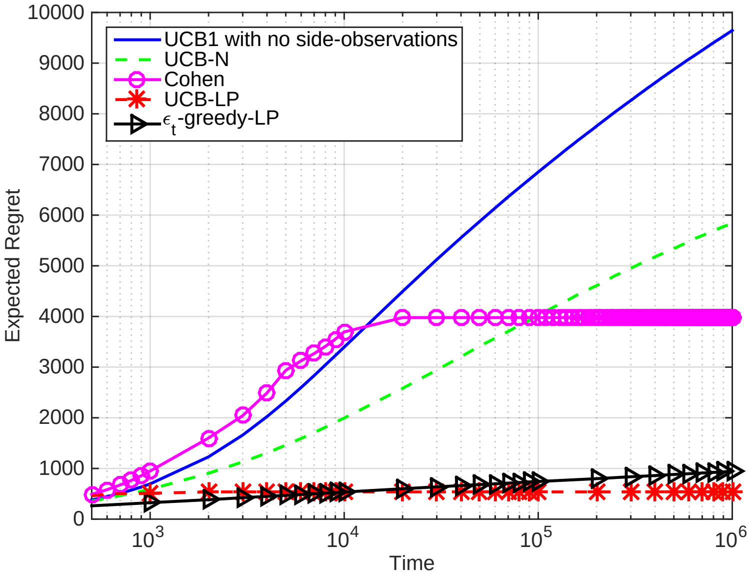

Figure 6: Regret comparison of all the policies for the routing problem

the UCB-N policy. The reason is that there is no non-trivial clique (clique with more than one element) in this problem due to the setting $\mathcal{K}_j = V_j$ and $\mathcal{K}_j \neq \mathcal{K}_a$ for any $j, a \in \mathcal{K}$. Intuitively, there does not exist two paths that can observe the outcome of each other. For the ϵ_t-greedy-LP policy, we let $c = 4$ and $d = 0.05$. From the regret performance, we observe that the improvement obtained by the UCB-N policy against the UCB1 policy is considerably greater than the results in Figure 3 and Figure 4. The reason behind this is that the bipartite graph in the routing problem is more dense and network size is small in the routing problem, which enables the UCB-N policy to take the advantage of side-observations. Overall, we see that our policies outperform the UCB-N policy and Cohen policy because our policies take the network structure into consideration, which enables us to trade off the exploration with exploitation more efficiently.

7.3 Asymptotic Behavior

We run a simulation to verify the result provided in Proposition 5. For each base-arm size N, we sequentially generate action j such that $e_{ij} = 1$ with probability p for any $i \in \mathcal{N}$ independently. Stopping time τ is the number of actions we have generated so that there are no useless base-arms in the network. Note that τ is an upper bound of $\gamma(G)$ and $\sum_{j\in\mathcal{K}} z_j^*$ as shown in Appendix B. Then, we solve the linear program P2 to obtain $\sum_{j\in\mathcal{K}} z_j^*$ and find a hitting set by a greedy algorithm.[4]

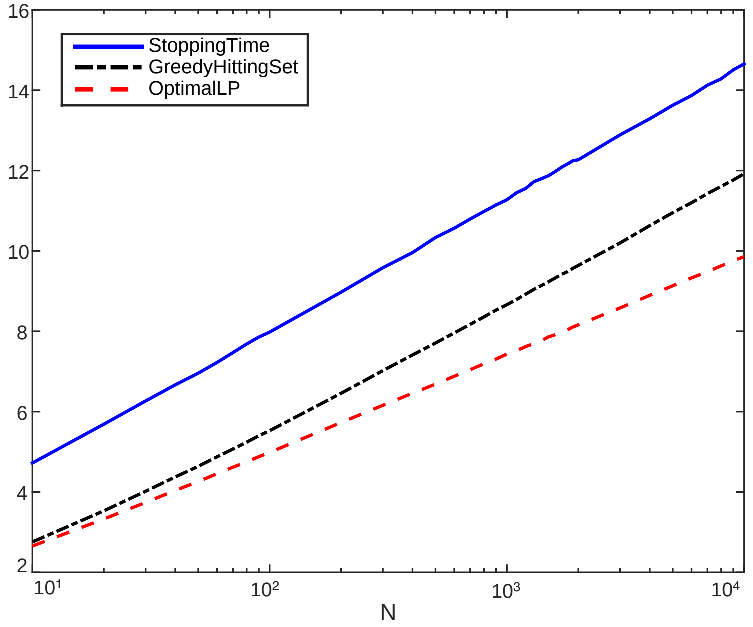

Figure 7: Erdos-Renyi random graph with p=0.5

Figure 7 shows the average result over 1000 samples for each N when $p = 0.5$. The numerical result verifies our theoretical result that $\sum_{j\in\mathcal{K}} z_j^*$ is upper-bounded by a logarithmic function of N asymptotically in Erdos-Renyi random graph. The reason why we are interested in the scaling order of $\sum_{j\in\mathcal{K}} z_j^*$ is that the traditional UCB1 policy suffers from the curse of dimensionality when applied in the real world, such as recommendation systems with thousands of items. However, our policies show a regret of $O(\sum_{j\in\mathcal{K}} z_j^* \log T)$, and $\sum_{j\in\mathcal{K}} z_j^*$ is upper-bounded by a logarithmic function of the number of unknowns, which makes our policies scalable in some large networks.

8. Summary

In this work, we introduce an important structural form of feedback available in many multiarmed bandits using the bipartite network structure. We obtained an asymptotic (with respect to time) lower bound as a function of the network structure on the regret of

[4] It is well known that hitting set problem is NP-complete. So we employ the greedy algorithm which brings in the node with the largest degree in the network during each iteration.

any uniformly good policy. Further, we proposed two policies: 1) the ϵ_t-greedy-LP policy, and 2) the UCB-LP policy, both of which are optimal in the sense that they achieve the asymptotic lower bound on the regret, up to a multiplicative constant that is independent of the network structure. These policies can have a better regret performance than existing policies for some important network structures. The ϵ_t-greedy-LP policy is a network-aware any-time policy, but its exploration is oblivious to the average rewards of the suboptimal actions. On the other hand, UCB-LP considers both the network structure and the average rewards of actions.

Important avenues of future work include the case of dynamic graphs – what would be the lower bound and corresponding algorithms if the graph structure remains known but changes with time? Recently Tossou et al. (2017) presented a novel extension of Thompson sampling algorithm for the setting of immediate neighbor feedback studied in Mannor and Shamir (2011); Caron et al. (2012); Buccapatnam et al. (2014). It would be interesting to see how to adapt Thompson sampling algorithm for the bipartite graph feedback structure introduced in our current work.

In what follows, we give the proofs of all propositions stated in the earlier sections. These proofs make use of Lemmas 15, 16, and 17, and Proposition 18 given in Section F.

Appendix A. Proof of Proposition 4

Let $\mathcal{U} = \{j : \mu_j < \mu^*\}$ be the set of suboptimal actions. Also, let $\Delta_j = \mu^* - \mu_j$. Also, $T_j(t)$ is the total number of times action j is chosen up to time t by policy $\boldsymbol{\phi}$. Let $M_i(t)$ be the total number of observations corresponding to base-arm i available at time t. From Proposition 18 given in the Appendix, we have,

$$\liminf_{t\to\infty} \frac{\mathbb{E}[M_i(t)]}{\log(t)} \geq \frac{1}{J_i(\theta_i)},\ \forall i \in \mathcal{N}. \tag{8}$$

An observation is received for base-arm i whenever any action in S_i is chosen. Hence,

$$M_i(t) = \sum_{j\in S_i} T_j(t). \tag{9}$$

Now, from Equations (8) and (9), for each $i \in \mathcal{N}$,

$$\liminf_{t\to\infty} \frac{\sum_{j\in S_i} \mathbb{E}[T_j(t)]}{\log(t)} \geq \frac{1}{J_i(\theta_i)}. \tag{10}$$

Using (10), we get the constraints of LP P_1. Further, we have from definition of regret that,

$$\liminf_{t\to\infty} \frac{R_\mu(t)}{\log(t)} = \liminf_{t\to\infty} \sum_{j\in\mathcal{U}} \Delta_j \frac{\mathbb{E}[T_j(t)]}{\log(t)}.$$

The above equation along with the constraints of the LP P_1 obtained from (10) gives us the required lower bound on regret.

Appendix B. Proof of Proposition 5

Here we consider a E-R random graph with each entry of the matrix E equals 1 with probability p, where $0 < p < 1$. Consider the following discrete stochastic process. χ_n are i.i.d., such that $\chi_n \subseteq [N]$ is sampled by the following steps: for each $i = 1, 2, .., N$, $i \in \chi_n$ with probability p. Let $q = 1 - p$. Then let τ be a stopping time defined as

$$\tau = \min\{n \geq 1, \cup_{j=1}^{n} \chi_j = [N]\} \tag{11}$$

The complement cdf of τ is the following.

$$P(\tau > n) = 1 - (1 - q^n)^N \tag{12}$$

1. Fix N. Given $0 < p < 1$, then $0 < q < 1$. Thus, $P(\tau = \infty) = 0$
2. What is the upper bound of $E(\tau)$?

$$(1 - q^n)^N > \exp(-\frac{q^n N}{1 - q^n}) \quad (since\ \ \ln(1 - x) > -\frac{x}{1 - x} \quad for \ \ 0 < x < 1) \tag{13}$$

Thus, we have

$$P(\tau > n) < 1 - \exp(-\frac{q^n N}{1-q^n}) \leq \frac{q^n N}{1-q^n} \quad (since\ 1 - e^x \leq -x) \tag{14}$$

Then we can bound the expectation of τ.

$$E(\tau) = \sum_{n=1}^{\infty} P(\tau > n) < \sum_{n=1}^{\infty} q^n \frac{N}{1-q} = \frac{qN}{(1-q)^2} \tag{15}$$

3. What is the upper bound of $P(\tau \leq n)$?

$$P(\tau \leq n) = (1-q^n)^N \leq \exp(-q^n N) \tag{16}$$

4. Does τ converge as N goes to infinity?
Given $\epsilon > 0$, as N goes to ∞,

$$P(\tau \leq (1-\epsilon)\log_{1/q} N) \leq \exp(-q^{(1-\epsilon)\log_{1/q} N} N) = \exp(-N^{\epsilon}) \to 0 \tag{17}$$

$$P(\tau > (1+\epsilon)\log_{1/q} N) \leq \frac{q^{(1+\epsilon)\log_{1/q} N} N}{1-q} = \frac{1}{(1-q)N^{\epsilon}} \to 0 \tag{18}$$

Since ϵ is arbitrary, we can have

$$P(\tau = \log_{1/q} N) \to 1 \quad as\ N \to \infty. \tag{19}$$

That is to say τ converges to $\log_{1/q} N$ in probability.

Suppose there are no useless base-arms in the network, i.e. $[K]$ is a hitting set. Then τ is less than K with probability 1. Given this information, $\gamma(G)$ should be upper bounded by $\log_{1/q} N$ as N goes to infinity.

Appendix C. Proof of Proposition 6

Let $(z_j^*)_{j \in \mathcal{K}}$ be the optimal solution of LP P_2. We will first prove the upper bound in Equation 2. Using the optimal solution $(w_j^*)_{j \in \mathcal{K}}$ of LP P_1, we construct a feasible solution satisfying constraints in LP P_2 in the following way: For actions $j \in \mathcal{K}$, let $z_j = \left(\max_{i \in \mathcal{N}} J_i(\theta_i)\right) w_j^*$. Then $(z_j)_{j \in \mathcal{K}}$ satisfy constraints for all base-arms $i \in \mathcal{N}$ because w_j^* satisfy constraints of LP P_1.

The feasible solution constructed in this way gives an upper bound on the optimal value of LP P_2. Hence,

$$\begin{aligned}
\sum_{j \in \mathcal{K}} z_j^* &\leq \sum_{j \in \mathcal{U}} z_j + |\mathcal{O}| \\
&\leq \sum_{j \in \mathcal{U}} \left(\max_{i \in \mathcal{N}} J_i(\theta_i)\right) w_j^* + |\mathcal{O}| \\
&\leq \frac{\max_{i \in \mathcal{N}} J_i(\theta_i)}{\min_{j \in \mathcal{U}} \Delta_j} \sum_{j \in \mathcal{U}} \Delta_j w_j^* + |\mathcal{O}| \\
&\leq \frac{\max_{i \in \mathcal{N}} J_i(\theta_i)}{\min_{j \in \mathcal{U}} \Delta_j} c_{\boldsymbol{\mu}} + |\mathcal{O}|
\end{aligned}$$

For the lower bound, any feasible solution of P_2, in particular $\mathbf{z}^*$, can be used to construct a feasible solution of P_1. For actions $j \in \mathcal{K}$, let $w_j = \dfrac{z_j^*}{\min_{i\in\mathcal{N}} J_i(\theta_i)}$. Then $(w_j)_{j\in\mathcal{K}}$ satisfies the constraints of LP P_1 and hence gives an upper bound on its optimal value. Therefore, we have

$$\begin{aligned}
c_{\boldsymbol{\mu}} &= \sum_{j\in\mathcal{U}} \Delta_j w_j^*, \\
&\leq \sum_{j\in\mathcal{K}} \frac{\Delta_j z_j^*}{\min_{i\in\mathcal{N}} J_i(\theta_i)} \\
&\leq \sum_{j\in\mathcal{K}} \frac{\max_{a\in\mathcal{U}} \Delta_a z_j^*}{\min_{i\in\mathcal{N}} J_i(\theta_i)}
\end{aligned}$$

which gives us the required lower bound.

Appendix D. Proof of Proposition 7

Since $\mathbf{z}^*$ satisfies the constraints in LP P_2, there is sufficient exploration within each suboptimal action's neighborhood. The proof is then a combination of this fact and the proof of Theorem 3 in Auer et al. (2002). Let $\bar{f}_j(t)$ be the random variable denoting the sample mean of all observations available for action j at time t. Let $\bar{f}^*(t)$ be the random variable denoting the sample mean of all observations available for an optimal action at time t. Fix a suboptimal action j. For some $\alpha > 1$, define m_i for each base-arm i as follows,

$$m_i = \frac{1}{\alpha} \frac{\sum_{j\in S_i} z_j^*}{\sum_{j\in\mathcal{K}} z_j^*} \sum_{m=1}^{t} \epsilon(m)$$

Let $\phi(t)$ be the action chosen by ϵ_t-greedy-LP policy at time t. The event $\{\phi(t) = j\}$ implies that either sampling a random action j for exploration or playing the best observed action j for exploitation. Then,

$$\mathbb{P}[\phi(t) = j] \leq \frac{\epsilon(t) z_j^*}{\sum_{a\in\mathcal{K}} z_a^*} + (1-\epsilon(t))\mathbb{P}[\bar{f}_j(t) \geq \bar{f}^*(t)]$$

The event $\{\bar{f}_j(t) \geq \bar{f}^*(t)\}$ implies that either $\{\bar{f}_j(t) \geq \mu_j + \frac{\Delta_j}{2}\}$ or $\{\bar{f}^*(t) \leq \mu^* - \frac{\Delta_j}{2}\}$ since $\mu_j + \frac{\Delta_j}{2} = \mu^* - \frac{\Delta_j}{2}$. We also have that,

$$\mathbb{P}[\bar{f}_j(t) \geq \bar{f}^*(t)] \leq \mathbb{P}\left[\bar{f}_j(t) \geq \mu_j + \frac{\Delta_j}{2}\right] + \mathbb{P}\left[\bar{f}^*(t) \leq \mu^* - \frac{\Delta_j}{2}\right].$$

The analysis of both the terms in the right hand side of the above expression is similar. Let $O_i^R(t)$ be the total number of observations available for base-arm i from the exploration iterations of the policy up to time t. Let $O_i(t)$ be the total number of observations available for base-arm i up to time t. By concentration inequalities, the probability that the empirical mean deviate from the expectation can be bounded given the number of observations. The number of observations for action j is lower-bounded by the number of observations from the exploration iterations. Hence, we have,

$$\begin{aligned}
\mathbb{P}\left[\bar{f}_j(t) \geq \mu_j + \frac{\Delta_j}{2}\right] &= \sum_{m=1}^{t} \mathbb{P}\left[\min_{i\in\mathcal{K}_j} O_i(t) = m; \bar{f}_j(t) \geq \mu_j + \frac{\Delta_j}{2}\right] \\
&= \sum_{m=1}^{t} \mathbb{P}\left[\bar{f}_j(t) \geq \mu_j + \frac{\Delta_j}{2} | \min_{i\in\mathcal{K}_j} O_i(t) = m\right] \mathbb{P}\left[\min_{i\in\mathcal{K}_j} O_i(t) = m\right] \\
&\leq \sum_{m=1}^{t} \mathbb{P}\left[\min_{i\in\mathcal{K}_j} O_i(t) = m\right] e^{\frac{-\Delta_j^2 m}{2}} \\
&\text{(follows from Chernoff-Hoeffding bound in Lemma 15)} \\
&\leq \sum_{m=1}^{t} \mathbb{P}\left[\min_{i\in\mathcal{K}_j} O_i^R(t) \leq m\right] e^{\frac{-\Delta_j^2 m}{2}} \\
&\leq \sum_{m=1}^{\lfloor m_0 \rfloor} \mathbb{P}\left[\min_{i\in\mathcal{K}_j} O_i^R(t) \leq m\right] + \sum_{m=\lfloor m_0 \rfloor+1}^{t} e^{\frac{-\Delta_j^2 m}{2}} \\
&\leq m_0 \mathbb{P}\left[\min_{i\in\mathcal{K}_j} O_i^R(t) \leq m_0\right] + \frac{2}{\Delta_j^2} e^{\frac{-\Delta_j^2 m_0}{2}} \\
&\qquad \left(\text{since } \sum_{m+1}^{\infty} e^{-ku} \leq \frac{1}{k} e^{-km}\right) \\
&\leq \sum_{i\in\mathcal{K}_j} m_0 \mathbb{P}\left[O_i^R(t) \leq m_0\right] + \frac{2}{\Delta_j^2} e^{\frac{-\Delta_j^2 m_0}{2}},
\end{aligned}$$

where $m_0 = \min_{i\in\mathcal{N}} m_i$.

Recall that $O_i^R(t)$ is the total number of observations for base-arm i from exploration. Now, we derive the bounds for the expectation and variance of $O_i^R(t)$ in order to use Bernstein's inequality.

$$\begin{aligned}
\mathbb{E}\left[O_i^R(t)\right] &= \sum_{m=1}^{t} \epsilon(m) \sum_{j\in S_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_j^*} \\
&= \frac{\sum_{j\in S_i} z_j^*}{\sum_{j\in\mathcal{K}} z_j^*} \sum_{m=1}^{t} \epsilon(m) = \alpha m_i \\
&\geq \alpha m_0
\end{aligned}$$

$$\begin{aligned}
var\left[O_i^R(t)\right] &= \sum_{m=1}^{t} \left(\epsilon(m) \sum_{j\in S_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_j^*}\right) \left(1 - \epsilon(m) \sum_{j\in S_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_j^*}\right) \\
&\leq \sum_{m=1}^{t} \epsilon(m) \sum_{j\in S_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_j^*} \\
&= \mathbb{E}[O_i^R(t)] = \alpha m_i
\end{aligned}$$

Now, using Bernstein's inequality given in Lemma 16, we have

$$\begin{aligned}
\mathbb{P}\left[O_i^R(t) \le m_0\right] &= \mathbb{P}\left[O_i^R(t) \le \mathbb{E}[O_i^R(t)] + m_0 - \alpha m_i\right] \\
&\le \mathbb{P}\left[O_i^R(t) \le \mathbb{E}[O_i^R(t)] + m_i - \alpha m_i\right] \\
&\le \exp\left(-\frac{(\alpha-1)^2 m_i^2}{2\alpha m_i + \frac{2}{3}(\alpha-1)m_i}\right) \\
&= \exp\left(-\frac{3(\alpha-1)^2}{8\alpha-2} m_i\right) = \exp(-r m_i)
\end{aligned}$$

where $r = \frac{3(\alpha-1)^2}{8\alpha-2}$. Now, we will obtain upper and lower bounds on m_i by plugging in the definition of $\epsilon(m)$. For the upper bound, for any $t > t' = \frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2}$,

$$\begin{aligned}
m_i &= \frac{\sum_{j\in S_i} z_j^*}{\alpha\sum_{j\in\mathcal{K}} z_j^*}\sum_{m=1}^{t}\epsilon(m) \\
&= \frac{\sum_{j\in S_i} z_j^*}{\alpha\sum_{j\in\mathcal{K}} z_j^*}t' + \frac{\sum_{j\in S_i} z_j^*}{\alpha\sum_{j\in\mathcal{K}} z_j^*}\sum_{m=t'+1}^{t}\frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2 m} \\
&\le \frac{c\delta}{\alpha d^2}\left(1 + \sum_{m=t'+1}^{t}\frac{1}{m}\right) \\
&\le \frac{c\delta}{\alpha d^2}\log\left(\frac{e^2 t}{t'}\right).
\end{aligned}$$

where $\delta = \max_{i\in\mathcal{N}}|S_i|$, denoting the maximum degree of the supports in the network. In the above, $\sum_{j\in S_i} z_j^* \le \delta$ because $z_j^* \le 1$, which is due to the fact that $\left(z_j^*\right)_{j\in\mathcal{K}}$ is the optimal solution of LP P_2. Next, for the lower bound, we use the fact that $\sum_{j\in S_i} z_j^* \ge 1$ for all i because $\left(z_j^*\right)_{j\in\mathcal{K}}$ satisfies the constraints of LP P_2. Thus

$$\begin{aligned}
m_i &\ge \frac{\sum_{j\in S_i} z_j^*}{\alpha\sum_{j\in\mathcal{K}} z_j^*}\sum_{m=t'+1}^{t}\frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2 m} \\
&\ge \frac{c}{\alpha d^2}\sum_{m=t'+1}^{t}\frac{1}{m} \\
&\ge \frac{c}{\alpha d^2}\log\left(\frac{t}{et'}\right).
\end{aligned}$$

Let $\lambda = \max_{j \in \mathcal{K}} |\mathcal{K}_j|$. Hence, combining the inequalities above,

$$\begin{aligned}
\mathbb{P}\left[\bar{f}_j(t) \geq \mu_j + \frac{\Delta_j}{2}\right] &\leq \sum_{i \in \mathcal{K}_j} m_0 \mathbb{P}\left[O_i^R(t) \leq m_0\right] + \frac{2}{\Delta_j^2} e^{-\frac{\Delta_j^2 m_0}{2}} \\
&\leq \sum_{i \in \mathcal{K}_j} m_0 \left(\frac{et'}{t}\right)^{cr/\alpha d^2} + \frac{2}{\Delta_j^2} e^{-\frac{\Delta_j^2 m_0}{2}} \\
&\leq \lambda \frac{c\delta}{\alpha d^2}\left(\log\left(\frac{e^2 t}{t'}\right)\right)\left(\frac{et'}{t}\right)^{cr/\alpha d^2} + \frac{2}{\Delta_j^2}\left(\frac{et'}{t}\right)^{\frac{c\Delta_j^2}{2\alpha d^2}}
\end{aligned}$$

Now, similarly for the optimal action, we have, for all $t > t'$

$$\mathbb{P}\left[\bar{f}^*(t) \leq \mu^* - \frac{\Delta_j}{2}\right] \leq \frac{\lambda c\delta}{\alpha d^2}\left(\frac{et'}{t}\right)^{cr/\alpha d^2} \log\left(\frac{e^2 t}{t'}\right) + \frac{2}{\Delta_j^2}\left(\frac{et'}{t}\right)^{\frac{c\Delta_j^2}{2\alpha d^2}}.$$

Combining everything, we have for any suboptimal action j, for all $t > t'$

$$\begin{aligned}
\mathbb{P}[\phi(t) = j] &\leq \frac{\epsilon(t) z_j^*}{\sum_{a \in \mathcal{K}} z_a^*} + (1 - \epsilon(t))\, P[\bar{f}_j(t) \geq \bar{f}^*(t)] \\
&\leq \frac{c z_j^*}{d^2 t} + P[\bar{f}_j(t) \geq \bar{f}^*(t)] \\
&\leq \frac{c z_j^*}{d^2 t} + \frac{2\lambda c\delta}{\alpha d^2}\left(\frac{et'}{t}\right)^{cr/\alpha d^2} \log\left(\frac{e^2 t}{t'}\right) + \frac{4}{\Delta_j^2}\left(\frac{et'}{t}\right)^{\frac{c\Delta_j^2}{2\alpha d^2}}
\end{aligned}$$

Appendix E. Proof of Proposition 10

The proof technique is similar to that in Auer and Ortner (2010). We will analyze the regret by conditioning on two disjoint events. The first event is that each suboptimal action a is eliminated by an optimal action on or before the first round m such that $\tilde{\Delta}_m < \Delta_a/2$. This happens with high probability and leads to logarithmic regret. The compliment of the first event yields linear regret in time but occurs with probability proportional to $1/T$. The main difference from the proof in Auer and Ortner (2010) is that on the first event, the number of times we choose each action j is proportional to $z_j^* \log(T)$ in the exploration iterations (i.e., when $|B_m| > 1$) of the policy. This gives us the required upper bound in terms of optimal solution $\mathbf{z}^*$ of LP P_2.

Let $*$ denote any optimal action. Let m^* denote the round in which the last optimal action $*$ is eliminated. For each suboptimal action j, define round $m_j := \min\{m : \tilde{\Delta}_m < \frac{\Delta_j}{2}\}$. For an optimal action j, $m_j = \infty$ by convention. Then, by the definition of m_j, for all rounds $m < m_j$, $\Delta_j \leq 2\tilde{\Delta}_m$, and

$$\frac{2}{\Delta_j} < 2^{m_j} = \frac{1}{\tilde{\Delta}_{m_j}} \leq \frac{4}{\Delta_j} < \frac{1}{\tilde{\Delta}_{m_j+1}} = 2^{m_j+1}. \tag{20}$$

From Lemma 17 in the Appendix, the probability that action j is not eliminated in round m_j by $*$ is at most $\frac{2}{T\tilde{\Delta}_{m_j}^2}$.

Let $I(t)$ be the action chosen at time t by the UCB-LP policy.
Let E_{m^*} be the event that all suboptimal actions with $m_j \leq m^*$ are eliminated by $*$ on or before their respective m_j. Then, the complement of E_{m^*}, denoted as $E^c_{m^*}$, is the event that there exists some suboptimal action j with $m_j \leq m^*$, which is not eliminated by round m_j. Let E^c_j be the event that action j is not eliminated by round m_j by $*$. Let $m_f = \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$ and $I(t)$ denote the action chosen at time t by the policy. Recall that regret is denoted by $R_{\boldsymbol{\mu}}(T)$. Let $\mathbb{P}[m^* = m]$ be denoted by p_m. Hence, $\sum_{m=0}^{m_f} p_m = 1$.

$$\begin{aligned}
\mathbb{E}\left[R_{\boldsymbol{\mu}}(T)\right] &= \sum_{m=0}^{m_f} \mathbb{E}\left[R_{\boldsymbol{\mu}}(T) | \{m^* = m\}\right] \mathbb{P}[m^* = m] \\
&= \sum_{m=0}^{m_f} \sum_{t=1}^{T} \sum_{j \in \mathcal{U}} \Delta_j \mathbb{P}\left[I(t) = j | \{m^* = m\}\right] p_m \\
&= \sum_{m=0}^{m_f} \sum_{t=1}^{T} \sum_{j \in \mathcal{U}} \Delta_j \mathbb{P}\left[\{I(t) = j\} \cap E_{m^*} | \{m^* = m\}\right] p_m \\
&\ + \sum_{m=0}^{m_f} \sum_{t=1}^{T} \sum_{j \in \mathcal{U}} \Delta_j \mathbb{P}\left[\{I(t) = j\} \cap E^c_{m^*} | \{m^* = m\}\right] p_m \\
&= (i) + (ii)
\end{aligned}$$

Next we will show that term (i) leads to logarithmic regret while term (ii) leads to a constant regret with time.
First, consider the term (ii) of the regret expression. For each $j \in \mathcal{U}$, we have,

$$\begin{aligned}
&\sum_{m=0}^{m_f} \sum_{t=1}^{T} \mathbb{P}\left[\{I(t) = j\} \cap E^c_{m^*} | \{m^* = m\}\right] \mathbb{P}[m^* = m] \\
&\qquad \leq \sum_{m=0}^{m_f} \sum_{t=1}^{T} \mathbb{P}\left[\{I(t) = j\} \cap \left(\cup_{a \in \mathcal{U}: m_a \leq m^*} E^c_a\right) | \{m^* = m\}\right] p_m \\
&\qquad \leq \sum_{m=0}^{m_f} \sum_{t=1}^{T} \Big(\mathbb{P}\left[\{I(t) = j\} | \left(\cup_{a \in \mathcal{U}: m_a \leq m^*} E^c_a\right), \{m^* = m\}\right] \\
&\qquad\qquad\qquad\qquad \mathbb{P}\left[\cup_{a \in \mathcal{U}: m_a \leq m^*} E^c_a | \{m^* = m\}\right] p_m \Big) \\
&\qquad \leq T \mathbb{P}\left[\cup_{a \in \mathcal{U}} E^c_a | \{m^* = m_f\}\right] \sum_{m=0}^{m_f} p_m \\
&\qquad \leq T \sum_{a \in \mathcal{U}} \frac{2}{T \tilde{\Delta}^2_{m_a}}, \\
&\qquad\quad \left(\text{using Lemma 17, } \mathbb{P}\left[E^c_a | \{m^* = m_f\}\right] \leq \frac{2}{T \tilde{\Delta}^2_{m_a}} \right) \\
&\qquad \leq \sum_{a \in \mathcal{U}} \frac{32}{\Delta^2_a},
\end{aligned}$$

where the last inequality follows from Equation (20). Hence, the term (ii) of regret is

$$\sum_{m=0}^{m_f}\sum_{t=1}^{T}\sum_{j\in\mathcal{U}}\Delta_j\mathbb{P}\left[\{I(t)=j\}\cap E^c_{m^*}|\{m^*=m\}\right]p_m$$
$$\leq \sum_{j\in\mathcal{U}}\Delta_j\sum_{a\in\mathcal{U}}\frac{32}{\Delta_a^2}=O(K^2). \tag{21}$$

Next, we consider the term (i). Recall that, in this term, we consider the case that all suboptimal actions j with $m_j \leq m^*$ are eliminated by $*$ on or before m_j.

$$\begin{aligned}
(i) &= \sum_{m=0}^{m_f}\sum_{t=1}^{T}\sum_{j\in\mathcal{U}}\Delta_j\mathbb{P}\left[\{I(t)=j\}\cap E_{m^*}|\{m^*=m\}\right]p_m\\
&= \sum_{m=0}^{m_f}\mathbb{E}\left[R_{\boldsymbol{\mu}}(T)|\{m^*=m\},E_{m^*}\right]\mathbb{P}[E_{m^*}|\{m^*=m\}]p_m\\
&\leq \sum_{m=0}^{m_f}\Big(\mathbb{E}\left[\text{Regret from }\{j:m_j\leq m^*\}|\{m^*=m\},E_{m^*}\right]\\
&\quad+\mathbb{E}\left[\text{Regret from }\{j:m_j> m^*\}|\{m^*=m\},E_{m^*}\right]\Big)p_m\\
&\leq \sum_{m=0}^{m_f}\Big(\mathbb{E}\left[\text{Regret from }\{j:m_j\leq m_f\}|\{m^*=m_f\},E_{m_f}\right]\\
&\quad+\mathbb{E}\left[\text{Regret from }\{j:m_j> m^*\}|\{m^*=m\},E_{m^*}\right]\Big)p_m\\
&\leq \mathbb{E}\left[R_{\boldsymbol{\mu}}(T)|\{m^*=m_f\},E_{m_f}\right]\sum_{m=0}^{m_f}p_m\\
&+\sum_{m=0}^{m_f}\mathbb{E}\left[\text{Regret from }\{j:m_j> m^*\}|\{m^*=m\},E_{m^*}\right]p_m\\
&=(ia)+(ib)
\end{aligned}$$

Once again, we will consider the above two terms separately. For the term (ia), under the event E_{m_f}, each suboptimal action j is eliminated by $*$ by round m_j. Define round $\bar{m}$ and the set B as follows:

$$\begin{aligned}
\bar{m} &= \min\{m:\sum_{j\in\mathcal{K}}z^*_j>\sum_{a:m_a>m}2^{-m+1}\},\\
B &= \{j\in\mathcal{U}:m_j>\bar{m}\}.
\end{aligned}$$

After round $\bar{m}$, Algorithm 2 chooses only those actions with $m_j > \bar{m}$. Also, by the definition of the **Reset** phase of Algorithm 2, we have that any suboptimal action $j\notin B$ is chosen (i.e. appears in the set A_m at round m) only until it is not in A_m or until $\bar{m}$, whichever happens first. Define $n_j = \min\{\bar{m}, \max\limits_{a:j\in G_a}\{m_a\}\}$ for each suboptimal action j, where $G_a = \bigcup_{i\in\mathcal{K}_a} S_i$ for action a. Then any suboptimal action $j \notin B$ is chosen for at most n_j rounds.

$$\begin{aligned}(ia) &= \mathbb{E}\left[R_{\boldsymbol{\mu}}(T)|\{m^* = m_f\}, E_{m_f}\right] \\ &\leq \sum_{j\in\mathcal{U}\backslash B} \Delta_j z_j^* \frac{2\log(T\tilde{\Delta}_{n_j}^2)}{\tilde{\Delta}_{n_j}^2} + \sum_{j\in B} \Delta_j \frac{2\log(T\tilde{\Delta}_{m_j}^2)}{\tilde{\Delta}_{m_j}^2} \\ &\leq \sum_{j\in\mathcal{U}\backslash B} \Delta_j z_j^* \frac{32\log(T\hat{\Delta}_j^2)}{\hat{\Delta}_j^2} + \sum_{j\in B} \Delta_j \frac{32\log(T\Delta_j^2)}{\Delta_j^2}, \qquad (22)\end{aligned}$$

where $\hat{\Delta}_j = \max\{2^{-\bar{m}+2}, \min_{a:j\in G_a}\{\Delta_a\}\}$ and (z_j^*) is the solution of LP P_2.
Finally, we consider the term (ib). Note that $T_j(m) \geq n(m), \forall j \in B_m, \forall m$. An optimal action $*$ is not eliminated in round m^* if (25) holds for $m = m^*$. Hence, using (26) and (27), the probability p_m that $*$ is eliminated by a suboptimal action in any round m^* is at most $\frac{2}{T\tilde{\Delta}_{m^*}^2}$. Hence, term (ib) is given as:

$$\begin{aligned}&\sum_{m=0}^{m_f} \mathbb{E}\left[\text{Regret from } \{j : m_j > m^*\}|\{m^* = m\}, E_{m^*}\right] p_m \\ &\quad\leq \sum_{m=0}^{m_f} \sum_{j\in\mathcal{U}:m_j\geq m} \frac{2}{T\tilde{\Delta}_m^2} . T \max_{a\in\mathcal{U}} \Delta_a \\ &\quad\leq \max_{a\in\mathcal{U}} \Delta_a \sum_{m=0}^{m_f} \sum_{j\in\mathcal{U}:m_j\geq m} \frac{2}{\tilde{\Delta}_m^2} \\ &\quad\leq \sum_{j\in\mathcal{U}} \sum_{m=0}^{m_j} \frac{2}{\tilde{\Delta}_m^2} \\ &\quad\leq \sum_{j\in\mathcal{U}} 2^{2m_j+2} \leq \sum_{j\in\mathcal{U}} \frac{64}{\Delta_j^2} = O(K). \qquad (23)\end{aligned}$$

Now we get the result (6) by combining the bounds in (21), (22), and (23).
Further, the definition of set B ensures that we have

$$\sum_{j\in B} \Delta_j \leq \sum_{j\in\mathcal{K}} z_j^*.$$

Also, using the Assumption 4, $\frac{32\Delta_j\log(T\hat{\Delta}_j^2)}{\hat{\Delta}_j^2}, \frac{32\log(T\Delta_j^2)}{\Delta_j^2}$ are bounded by $C\log(T)$, where $C = \frac{32}{\min_{j\in\mathcal{U}}\Delta_j^2}$, is a constant independent of network structure. When one checks the feasibility of C, note that $\hat{\Delta}_j \geq \min_{a:j\in G_a}\Delta_a$ by definition and $\Delta_j \leq 1$ for any j since the rewards are bounded by 1. Hence, (22) can be bounded as:

$$\begin{aligned}\sum_{j\in\mathcal{U}\setminus B}\Delta_j z_j^* \frac{32\log(T\hat{\Delta}_j^2)}{\hat{\Delta}_j^2} &+ \sum_{j\in B}\Delta_j \frac{32\log(T\Delta_j^2)}{\Delta_j^2}\\ &\le \sum_{j\in\mathcal{U}\setminus B} z_j^* C\log(T) + \sum_{j\in B}\Delta_j C\log(T)\\ &\le \sum_{j\in\mathcal{U}\setminus B} z_j^* C\log(T) + \sum_{j\in B} 2^{-\bar{m}+1} C\log(T)\\ &\le 2\sum_{j\in\mathcal{K}} z_j^* C\log(T). \qquad (24)\end{aligned}$$

Hence, we get (7) from (24), (21), and (23).

Appendix F. Supplementary Material

$S_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ denotes the sample mean of the random variables $X_1, \ldots, X_n$. The first two lemmas below state the Chernoff-Hoeffding inequality and Bernstein's inequality.

Lemma 15 *Let $X_1, \ldots, X_n$ be a sequence of random variables with support $[0, 1]$ and $\mathbb{E}[X_t] = \mu$ for all $t \le n$. Let $S_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. Then, for all $\epsilon > 0$, we have,*

$$\mathbb{P}[S_n \ge \mu + \epsilon] \le e^{-2n\epsilon^2}$$
$$\mathbb{P}[S_n \le \mu - \epsilon] \le e^{-2n\epsilon^2}.$$

Lemma 16 *Let $X_1, \ldots, X_n$ be a sequence of random variables with support $[0, 1]$ and $\sum_{k=1}^{t} var[X_k|X_1, \ldots, X_{k-1}] \le \sigma^2$ for all $t \le n$. Let $S_n = \sum_{j=1}^{n} X_j$. Then, for all $\epsilon > 0$, we have,*

$$\mathbb{P}[S_n \ge \mathbb{E}[S_n] + \epsilon] \le \exp\left\{-\frac{\epsilon^2}{2\sigma^2 + \frac{2}{3}\epsilon}\right\}$$
$$\mathbb{P}[S_n \le \mathbb{E}[S_n] - \epsilon] \le \exp\left\{-\frac{\epsilon^2}{2\sigma^2 + \frac{2}{3}\epsilon}\right\}.$$

The next lemma is used in the proof of Proposition 10.

Lemma 17 *The probability that action j is not eliminated in round m_j by $*$ is at most $\frac{2}{T\tilde{\Delta}_{m_j}^2}$.*

Proof Let $\bar{f}_j(m)$ be the sample mean of all observations for action j available in round m. Let $\bar{f}^*(m)$ be the sample mean of the optimal action. The constraints of LP P_2 ensure that at the end of each round m, for all actions in B_m, we have at least $n(m) := \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$ observations. The reason is as follows. The set A_m contains set B_m. In particular, $A_m = \cup_{i\in D_m} S_i$ and $D_m = \cup_{j\in B_m} \mathcal{K}_j$. If each action j in A_m is played z_j^* times, then all the base-arms in D_m have at least 1 observations according the constraints of LP P_2. Thus, the actions in B_m have at least 1 observations. In sum, for all actions in B_m, we have at least $n(m) - n(m-1)$ observations at round m. Thus, we have at least $n(m)$ observations for all actions in B_m.

Now, for $m = m_j$, if we have,

$$\bar{f}_j(m) \leq \mu_j + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n(m)}} \quad \text{and} \quad \bar{f}^*(m) \geq \mu^* - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n(m)}}, \tag{25}$$

then, action j is eliminated by $*$ in round m_j. In fact, in round m_j, we have

$$\sqrt{\frac{\log(T\tilde{\Delta}_{m_j}^2)}{2n(m_j)}} \leq \frac{\tilde{\Delta}_{m_j}}{2} < \frac{\Delta_j}{4}.$$

Hence, in the elimination phase of the UCB-LP policy, if (25) holds for action j in round m_j, we have,

$$\begin{aligned}
\bar{f}_j(m_j) + \sqrt{\frac{\log(T\tilde{\Delta}_{m_j}^2)}{2n(m_j)}} &\leq \mu_j + 2\sqrt{\frac{\log(T\tilde{\Delta}_{m_j}^2)}{2n(m_j)}} \\
&< \mu_j + \Delta_j - 2\sqrt{\frac{\log(T\tilde{\Delta}_{m_j}^2)}{2n(m_j)}} \\
&= \mu^* - 2\sqrt{\frac{\log(T\tilde{\Delta}_{m_j}^2)}{2n(m_j)}} \\
&\leq \bar{f}^*(m_j) - \sqrt{\frac{\log(T\tilde{\Delta}_{m_j}^2)}{2n(m_j)}},
\end{aligned}$$

and action j is eliminated. Hence, the probability that action j is not eliminated in round m_j is the probability that either one of the inequalities in (25) do not hold. Using Chernoff-Hoeffding bound (Lemma 15), we can bound this as follows,

$$\mathbb{P}\left[\bar{f}_j(m) > \mu_j + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n(m)}}\right] \leq \frac{1}{T\tilde{\Delta}_m^2} \tag{26}$$

$$\mathbb{P}\left[\bar{f}^*(m) < \mu^* - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n(m)}}\right] \leq \frac{1}{T\tilde{\Delta}_m^2}. \tag{27}$$

Summing the above two inequalities for $m = m_j$ gives us that the probability that action j is not eliminated in round m_j by $*$ is at most $\frac{2}{T\tilde{\Delta}_{m_j}^2}$. ■

The next proposition is a modified version of Theorem 2 in Lai and Robbins (1985). We use it to obtain the regret lower bound in Proposition 4.

Proposition 18 *Suppose Assumptions 1, 2, and 3 hold. Let $M_i(t)$ be the total number of observations for such a base-arm i, for which $\vec{\theta} \in \Theta_i$. Then, under any uniformly good policy $\boldsymbol{\phi}$, we have that*

$$\liminf_{t\to\infty} \frac{\mathbb{E}[M_i(t)]}{\log(t)} \geq \frac{1}{J_i(\theta_i)}.$$

Proof By definition of $J_i(\theta_i)$, for every $\epsilon > 0$, there exists a $\theta_i' \in \mathcal{B}_i(\theta_i)$ such that $J_i(\theta_i) < D(\theta_i||\theta_i') < (1+\epsilon)J_i(\theta_i)$.

Now, under $\vec{\theta}_i' = [\theta_1, \dots, \theta_i', \dots \theta_N]$, there exists an action $k \in \mathcal{S}_i$ such that k is the unique optimal action. Then, for any uniformly good policy, for $0 < b < \delta$,

$$\mathbb{E}_{\vec{\theta}_i'}[t - T_k(t)] = o(t^b)$$

and therefore,

$$\mathbb{P}_{\vec{\theta}_i'}\left[T_k(t) < (1-\delta)\log(t)/D(\theta_i||\theta_i')\right] = o(t^{b-1}),$$

similar to the asymptotic lower bound proof in Lai and Robbins (1985).

Let $M_i(t)$ be the total number of observations for base-arm i. Then $M_i(t) \geq T_k(t)$, since choosing any action in $\mathcal{S}_i$ gives observations for i. Hence,

$$\mathbb{P}_{\vec{\theta}_i'}\left[M_i(t) < (1-\delta)\log(t)/D(\theta_i||\theta_i')\right] = o(t^{b-1}),$$

Now the rest of the proof of Theorem 2 in Lai and Robbins (1985) applies directly to $M_i(t)$. We will repeat it below for completeness. Let $(Y_i(r))_{r\geq 1}$ be the observations drawn from distribution F_i and define

$$L_m = \sum_{r=1}^{m} \log\left(\frac{g(Y_i(r);\theta_i)}{g(Y_i(r);\theta_i')}\right).$$

Now, we have that $\mathbb{P}_{\vec{\theta}_i'}[C_t] = o(t^{b-1})$ where $C_t = \{M_i(t) < (1-\delta)\log(t)/D(\theta_i||\theta_i')$ and $L_{M_i(t)} \leq (1-b)\log(t)\}$.

Now, we use the change of measure arguments.

$$\mathbb{P}_{\vec{\theta}_i'}[M_1(t) = m_1, \dots, M_N(t) = m_N, L_{m_i} \leq (1-b)\log(t)] \tag{28}$$

$$= \int_{\{M_1(t)=m_1,\dots,M_N(t)=m_N, L_{m_i}\leq(1-b)\log(t)\}} \Pi_{r=1}^{m_i} \frac{g(Y_i(r);\theta_i')}{g(Y_i(r);\theta_i)} dP_{\vec{\theta}_i} \tag{29}$$

$$\geq \exp(-(1-b)\log(t))\mathbb{P}_{\vec{\theta}_i}[M_1(t) = m_1, \dots, M_N(t) = m_N, L_{m_i} \leq (1-b)\log(t)] \tag{30}$$

Since C_t is a disjoint union of events of the form $\{M_1(t) = m_1, \dots, M_N(t) = m_N, L_{m_i} \leq (1-b)\log(t)\}$ with $m_i < (1-\delta)\log(t)/D(\theta_i||\theta_i')$, it follows that

$$\mathbb{P}_{\vec{\theta}}[C_t] \leq t^{1-b}\mathbb{P}_{\vec{\theta}_i'}[C_t] \to 0.$$

So far, we show that the probability of the event C_t goes to 0 as t goes to infinity. If we show the event $\{L_{M_i(t)} \leq (1-b)\log(t)|M_i(t) < (1-\delta)\log(t)/D(\theta_i||\theta_i')\}$ occurs almost surely, then we show the probability of $\{M_i(t) < (1-\delta)\log(t)/D(\theta_i||\theta_i')\}$ goes to 0 as t goes to infinity, which is the desired result. By strong law of large numbers $L_m/m \to D(\theta_i||\theta_i')$ as $m \to \infty$ and $\max_{r\leq m} L_r/m \to D(\theta_i||\theta_i')$ almost surely. Now, since $1-b > 1-\delta$, it follows that as $t \to \infty$,

$$\mathbb{P}_{\vec{\theta}}\left[L_r > (1-b)\log(t) \text{ for some } r < (1-\delta)\log(t)/D(\theta_i||\theta_i')\right] \to 0. \tag{31}$$

Hence, we have that as $t \to \infty$,

$$\mathbb{P}_{\vec{\theta}}\left[M_i(t) < (1-\delta)\log(t)/D(\theta_i||\theta_i')\right] \to 0.$$

By choosing ϵ, δ appropriately, this translates to

$$\liminf_{t\to\infty} \frac{\mathbb{E}[M_i(t)]}{\log(t)} \geq \frac{1}{J_i(\theta_i)}.$$

■

Proposition 19 *When the horizon is unknown, start the UCB-LP algorithm with $\tilde{T}_0 = 2$ and increase $\tilde{T}$ after reaching $\tilde{T}$ steps by setting $\tilde{T}_{l+1} = \tilde{T}_l^2$. The regret of unknown horizon UCB-LP is bounded by*

$$\sum_{j\in\mathcal{U}\setminus B} \frac{64\Delta_j z_j^*}{\hat{\Delta}_j^2} \log(T\hat{\Delta}_j^2) + \sum_{j\in B} \frac{64\log(T\Delta_j^2)}{\Delta_j} + O(K^2 \log_2\log_2 T). \tag{32}$$

Proof When the horizon is unknown, start the UCB-LP algorithm with $\tilde{T}_0 = 2$ and increase $\tilde{T}$ after reaching $\tilde{T}$ steps by setting $\tilde{T}_{l+1} = \tilde{T}_l^2$. Thus, $\tilde{T}_l = 2^{2^l}$ until reaching horizon T. Also, the period in which horizon is reached is denoted by L. Note that $2 \leq L \leq \log_2\log_2 T$.

In any period $l, (0 \leq l \leq L)$, UCB-LP uses $\tilde{T}_l$ as input. Note that $\bar{m}, m_j, B$ and $\hat{\Delta}_j$ are independent of $\tilde{T}_l$, thus l. Recall that regret of UCB-LP is bounded by (6). The regret of UCB-LP with unknown horizon is upper bounded by the summation over all the periods.

$$\sum_{l=0}^{L}\left[\sum_{j\in\mathcal{U}\setminus B} \Delta_j z_j^* \frac{32\log(\tilde{T}_l\hat{\Delta}_j^2)}{\hat{\Delta}_j^2} + \sum_{j\in B} \frac{32\log(\tilde{T}_l\Delta_j^2)}{\Delta_j} + O(K^2)\right] = (i) + (ii) + (iii).$$

First, we consider the term (i). We can plug in the definition of $\tilde{T}_l$ into (i).

$$\begin{aligned}
(i) &= \sum_{l=0}^{L}\sum_{j\in\mathcal{U}\setminus B} \Delta_j z_j^* \frac{32\log(\tilde{T}_l\hat{\Delta}_j^2)}{\hat{\Delta}_j^2} \\
&= \sum_{j\in\mathcal{U}\setminus B} \frac{32\Delta_j z_j^*}{\hat{\Delta}_j^2} \sum_{l=0}^{L} \log(2^{2^l}\hat{\Delta}_j^2) \\
&= \sum_{j\in\mathcal{U}\setminus B} \frac{32\Delta_j z_j^*}{\hat{\Delta}_j^2} \left((\log 2)\sum_{l=0}^{L} 2^l + (L+1)\log\hat{\Delta}_j^2\right) \\
&\leq \sum_{j\in\mathcal{U}\setminus B} \frac{32\Delta_j z_j^*}{\hat{\Delta}_j^2} \left(2^{L+1}(\log 2) + (L+1)\log\hat{\Delta}_j^2\right) \\
&\leq \sum_{j\in\mathcal{U}\setminus B} \frac{32\Delta_j z_j^*}{\hat{\Delta}_j^2} \left(2\log T + (L+1)\log\hat{\Delta}_j^2\right) \quad (\textit{since } L \leq \log_2\log_2 T) \\
&\leq \sum_{j\in\mathcal{U}\setminus B} \frac{64\Delta_j z_j^*}{\hat{\Delta}_j^2} \log(T\hat{\Delta}_j^2) \quad (\textit{since } (L+1)\log\hat{\Delta}_j^2 \leq 2\log\hat{\Delta}_j^2)
\end{aligned}$$

Similarly, we have that $(ii) \leq \sum_{j \in B} \frac{64 \log(T\Delta_j^2)}{\Delta_j}$. Now, we directly sum up the bound for term (iii).

$$(iii) \leq \sum_{l=0}^{L} O(K^2) \leq (L+1)O(K^2) = O(K^2 \log_2 \log_2 T).$$

Hence, by combining the results above, the regret of unknown horizon is bounded by

$$\sum_{j \in \mathcal{U} \setminus B} \frac{64\Delta_j z_j^*}{\hat{\Delta}_j^2} \log(T\hat{\Delta}_j^2) + \sum_{j \in B} \frac{64 \log(T\Delta_j^2)}{\Delta_j} + O(K^2 \log_2 \log_2 T).$$

■

References

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, May 2002. ISSN 0885-6125.

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

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. $\mathcal{X}$-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. *SIGMETRICS Perform. Eval. Rev.*, 42(1):289–300, June 2014. ISSN 0163-5999.

S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. In *UAI*, pages 142–151. AUAI Press, 2012.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. *CoRR*, abs/1605.07018, 2016.

Colin Cooper, Ralf Klasing, and Michele Zito. Lower bounds and algorithms for dominating sets in web graphs. *Internet Mathematics*, 2:275–300, 2005.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.

M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 135–142. ACM, 2010.

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

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670. ACM, 2010.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *NIPS*, pages 684–692, 2011.

Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal. Multi-armed bandit problems with dependent arms. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 721–728, New York, NY, USA, 2007. ACM.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, 2010.

Aristide C. Y. Tossou, Christos Dimitrakakis, and Devdatt Dubhashi. Thompson sampling for stochastic bandits with graph feedback. *CoRR*, abs/ 1701.04238, 2017.

Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.