Stochastic Graphical Bandits with Adversarial Corruptions

Shiyin Lu, Guanghui Wang, Lijun Zhang*
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{lusy, wanggh, zhanglj}@lamda.nju.edu.cn

Abstract

We study bandits with graph-structured feedback, where a learner repeatedly selects an arm and then observes rewards of the chosen arm as well as its neighbors in the feedback graph. Existing work on graphical bandits assumes either stochastic rewards or adversarial rewards, both of which are extremes and appear rarely in real-world scenarios. In this paper, we study graphical bandits with a reward model that interpolates between the two extremes, where the rewards are overall stochastically generated but a small fraction of them can be adversarially corrupted. For this problem, we propose an online algorithm that can utilize the stochastic pattern and also tolerate the adversarial corruptions. The main idea is to restrict exploration to carefully-designed independent sets of the feedback graph and perform exploitation by adopting a soft version of arm elimination. Theoretical analysis shows that our algorithm attains an $O(\alpha \ln K \ln T + \alpha C)$ regret, where α is the independence number of the feedback graph, K is the number of arms, T is the time horizon, and C quantifies the total corruptions introduced by the adversary. The effectiveness of our algorithm is demonstrated by numerical experiments.

Introduction

As a powerful sequential decision-making model, the multi-armed bandits (MAB) have found applications in a variety of real-world scenarios, such as cell planning (Maghsudi and Hossain 2016), medical trials (Villar, Bowden, and Wason 2015), and online advertising (Schwartz, Bradlow, and Fader 2017). In MAB, learning proceeds in a sequence of consecutive rounds. At each round, a learner first chooses one of K arms to play and then receives a reward of the chosen arm. The learner's goal is to maximize the total rewards over T rounds, which requires carefully balancing the trade-off between exploration (trying different arms to avoid missing the optimal arm) and exploitation (sticking to the empirically best arm to accumulate more rewards). The standard metric for MAB is the regret, which is the cumulative rewards obtained by always selecting the optimal arm in hindsight minus the total rewards received by the learner. Over the past decades, regret theories and effective algorithms for MAB have been well developed (Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2020).

A natural extension of MAB is the graphical bandits (GB), initially introduced by Mannor and Shamir (2011). In GB, there exists a undirected feedback graph G with nodes corresponding to K arms and edges characterizing the feedback structure. Specifically, an edge (u, v) in graph G indicates that whenever the learner plays arm u or arm v, the rewards of both u and v are revealed to the learner. Compared to MAB, the main advantage of GB is that it captures the side-information about rewards of unselected arms. Such side-information widely exist in real-world applications (Alon et al. 2017). For example, in online advertising, if two ads promote similar products, they can be connected with an edge in the feedback graph: When a user clicks one ad, it is likely that the user also takes interest in the other ad. In social networks, two persons following each other often share similar preferences, so behaviors of one person also reveal information on the other. Furthermore, from a theoretical view, GB generalizes and unifies two popular online learning paradigms: MAB and PEA (Prediction with Expert Advice, Cesa-Bianchi and Lugosi, 2006) in the sense that GB reduces to MAB and PEA when the feedback graph is empty and complete, respectively.

Since the seminal work of Mannor and Shamir (2011), there has been a large body of research on GB, which can be divided into two categories according to the reward model. One category is the stochastic GB (Caron et al. 2012; Buccapatnam, Eryilmaz, and Shroff 2014; Cohen, Hazan, and Koren 2016; Tossou, Dimitrakakis, and Dubhashi 2017; Liu, Zheng, and Shroff 2018; Liu, Buccapatnam, and Shroff 2018; Hu, Mehta, and Pan 2019; Lykouris, Tardos, and Wali 2020), where the reward of each arm is stochastically generated. The other is the adversarial GB (Mannor and Shamir 2011; Alon et al. 2013; Kocák et al. 2014; Alon et al. 2015; Neu 2015; Kocák, Neu, and Valko 2016; Lykouris, Sridharan, and Tardos 2018; Feng and Loh 2018; Rangi and Franceschetti 2019; Arora, Marinov, and Mohri 2019; Lee, Luo, and Zhang 2020), in which rewards of arms are determined by an adversary and can be hence nearly arbitrary. While the stochastic GB can achieve a logarithmic regret bound, it is too optimistic and the stochastic assumption on reward is too stringent. By contrast, the adversarial GB has wider applicability, but it is too pessimistic and only enjoys

*Lijun Zhang is the corresponding author.

a regret bound that scales with $\sqrt{T}$. Thus, a natural question arises: Is there a bandits model that lies between the stochastic and adversary worlds and admits regret guarantees only slightly worsen than the logarithmic regret bound? In fact, this question has been answered affirmatively in the context of MAB (Seldin and Slivkins 2014; Lykouris, Mirrokni, and Paes Leme 2018; Zimmert and Seldin 2019; Gupta, Koren, and Talwar 2019) and PEA (Amir et al. 2020), but still remains open for GB.

In this paper, we provide an affirmative answer to this question for GB. Inspired by previous work (Lykouris, Mirrokni, and Paes Leme 2018; Gupta, Koren, and Talwar 2019), we study the adversarially-corrupted stochastic setting and formulate a new graphical bandits model called stochastic graphical bandits with adversarial corruptions, where rewards of arms at each round are first drawn from some unknown distributions and then can be corrupted by an adversary with a total corruption budget. As a motivating example, consider a search engine that offers pay-per-click advertising services. When a user comes in, the search engine chooses an ad to display and obtains a reward if the ad is clicked. While for an ad most feedback from users (i.e., click or not) follows a stochastic pattern, some feedback may be maliciously simulated by botnets and get corrupted, resulting in the notorious phenomenon of click fraud (Wilbur and Zhu 2009).

For stochastic graphical bandits with adversarial corruptions, we propose an online algorithm that can exploit the generally stochastic nature of rewards and be also robust to the reward corruptions introduced by the adversary. Our algorithm builds on the BARBAR method (Gupta, Koren, and Talwar 2019), which is designed for stochastic MAB with adversarial corruptions. We extend BARBAR to the graphical bandits setting and leverage the graph-structured feedback to reduce the regret suffered from exploration. Specifically, we construct carefully-designed independent sets of the feedback graph and restrict the exploration to these independent sets instead of exploring on the whole arm set. Theoretical analysis shows that our algorithm enjoys an $O(\alpha \ln K \ln T + \alpha C)$ regret bound, where α is the independence number of the feedback graph and C is the amount of corruption. Compared to the $O(K \ln K \ln T + KC)$ regret bound of BARBAR, our result implies that in the corrupted setting, replacing pure bandit feedback with graph-structured feedback can also reduce the regret by a K/α factor, which is consistent with the existing literature on non-corrupted graphical bandits.

Besides extending BARBAR to graphical bandits, we also make improvements and refinements of BARBAR, including a different estimation of mean reward and a smaller constant for epoch length, which help to reduce the leading constant factor in the regret bound of BARBAR by more than 500 times and achieve better empirical performance. Finally, for stochastic PEA with adversarial corruptions, a special case of our setting with $\alpha = 1$, the existing algorithm only enjoys an expected regret bound and requires the optimal expert to be unique (Amir et al. 2020). By contrast, our algorithm achieves a high probability regret bound and can handle scenarios with multiple optimal experts. Finally, we conduct numerical experiments to demonstrate the effectiveness of our algorithm.

Related Work

In this section, we briefly review the related work.

Stochastic Graphical Bandits

The study of stochastic graphical bandits was initiated by Caron et al. (2012), who proposed an elegant extension of UCB termed UCB-N, where the estimated mean rewards are updated for not only the chosen arm but also its neighbors. For this algorithm, Caron et al. (2012) derived an $O(\bar{\chi} \ln T + K)$ regret bound, where $\bar{\chi}$ is the clique covering number of the feedback graph. In a recent work, Hu, Mehta, and Pan (2019) proposed a variant of UCB-N called UCB-NE, where extra exploration was introduced so as to improve the regret bound to $O(\bar{\chi} \ln T)$.

Instead of following UCB, Buccapatnam, Eryilmaz, and Shroff (2014) proposed a successive elimination method, which, though, is termed as UCB-LP. By leveraging graph feedback to adjust the exploration rate for each arm, UCB-LP attains an $O\big(\gamma \ln T + KD\big)$ regret, where γ is the domination number of the feedback graph G and D is the maximum degree in G. In a subsequent work, Cohen, Hazan, and Koren (2016) considered a harder setting where the feedback graph is directed, time-variant, and not fully revealed to the learner. Cohen, Hazan, and Koren (2016) also developed an elimination-based algorithm, which enjoys an $O\big(\alpha_{\max} \ln K \ln (KT)\big)$ regret bound, where $\alpha_{\max}$ is the maximum independence number of the feedback graph over T rounds.

Another popular technique to handle the exploration and exploitation dilemma inherent in any stochastic bandits problem is Thompson Sampling (TS, Thompson, 1933). For stochastic graphical bandits, Tossou, Dimitrakakis, and Dubhashi (2017) and Liu, Buccapatnam, and Shroff (2018) proposed several variants of TS with $O(\sqrt{\bar{\chi} T})$ Bayesian regret bounds. A refined Bayesian regret bound of $O(\sqrt{\alpha T})$ was later provided by Liu, Zheng, and Shroff (2018). In a subsequent work, Hu, Mehta, and Pan (2019) proved an $O(\bar{\chi} \ln T)$ frequentist regret bound for the TS-N algorithm developed by Tossou, Dimitrakakis, and Dubhashi (2017). Very recently, Lykouris, Tardos, and Wali (2020) proposed a novel layering technique and derived an improved $O(\alpha \ln^2 T)$ frequentist regret bound for both UCB-N and TS-N.

Stochastic Learning with Adversarial Corruptions

In their seminal work, Lykouris, Mirrokni, and Paes Leme (2018) introduced a new bandits model termed as stochastic MAB with adversarial corruptions. For this model, they first showed that with prior knowledge of the amount of corruption, the active arm elimination method (AAE, Even-Dar, Mannor, and Mansour, 2006) with enlarged confidence intervals suffices to guarantee logarithmic regrets. Then, they proposed a multi-layer technique which enables AAE to adapt to unknown amount of corruption and attain an $O(KC \ln K \ln T)$ regret. Later, Gupta, Koren, and Talwar

(2019) improved this bound to $O(K \ln K \ln T + KC)$ by developing a novel variant of AAE, which gradually reduces the chance of being selected for arms with bad empirical performance instead of permanently eliminating them. Gupta, Koren, and Talwar (2019) also established an $\Omega(K \ln T + C)$ lower bound indicating a gap of a factor K between the lower and the upper bounds. This gap was partially bridged by Zimmert and Seldin (2019), who proved that online mirror descent with Tsallis-INF regularizer can achieve the optimal $O(K \ln T + C)$ bound in expectation, provided that the optimal arm is unique. Recently, the adversarially-corrupted stochastic reward model is extended to prediction with expert advice (Amir et al. 2020), assortment optimization (Chen, Krishnamurthy, and Wang 2019), Gaussian bandits (Bogunovic, Krause, and Scarlett 2020), linear bandits (Kapoor, Patel, and Kar 2019; Li, Lou, and Shan 2019), and reinforcement learning (Lykouris et al. 2019). Instead of studying the budget-bounded corruption setting, several papers focus on the scenario where the rewards are corrupted with a fixed probability (Altschuler, Brunel, and Malek 2019; Kapoor, Patel, and Kar 2019; Guan et al. 2020).

While the above work studies the regime between the stochastic and the adversarial worlds, there has also been a surge of research interest in developing algorithms with optimal regret bounds for both worlds (Bubeck and Slivkins 2012; Seldin and Slivkins 2014; Auer and Chiang 2016; Seldin and Lugosi 2017; Wei and Luo 2018; Zimmert and Seldin 2019; Zimmert, Luo, and Wei 2019; Mourtada and Gaïffas 2019). Finally, there exists an orthogonal line of research that investigates attack strategies against stochastic bandits algorithms (Jun et al. 2018; Ma et al. 2018; Liu and Shroff 2019; Liu and Lai 2020; Garcelon et al. 2020).

Problem Setup

We study stochastic graphical bandits with adversarial corruptions. Let $[K] = \{1, \ldots, K\}$ be the arm set and $G = (V, E)$ be the feedback graph with $V = [K]$ and $E \subseteq V \times V$. For an arm $a \in [K]$, we denote by $\phi(a)$ the set comprised of a and its neighbors

$$\phi(a) = \{a\} \cup \big\{a' \in [K] \mid (a', a) \in E\big\}.$$

There is a learner and an adversary interacting with each other over T rounds. In each round t,

(1) A stochastic reward for each arm $r_t(a) \in [0, 1]$ is generated according to its reward distribution.

(2) The adversary observes the stochastic rewards of all arms $\big\{r_t(a)\big\}_{a \in [K]}$ and then determines the corrupted reward $\tilde{r}_t(a) \in [0, 1]$ for each arm $a \in [K]$.

(3) The learner chooses an arm $I_t \in [K]$ and then receives the corrupted reward of the chosen arm $\tilde{r}_t(I_t)$ and additionally observes the corrupted reward $\tilde{r}_t(a)$ of each arm a that is adjacent to the chosen arm $(I_t, a) \in E$.

It is worth pointing out that the above protocol implies an adaptive adversary: The corrupted rewards in round t can depend on the stochastic rewards up to round t as well as the corrupted rewards and the learner's choices before round t.

Following Gupta, Koren, and Talwar (2019), we use pseudo-regret, or simply regret, to evaluate the learner's performance. Let $\mu(a)$ denote the expectation of the reward distribution of arm $a \in [K]$ and $a^* \in \arg\max_{a \in [K]} \mu(a)$ be an optimal arm. The regret is defined as

$$\mathrm{R}(T) = \sum_{t=1}^{T} \mu(a^*) - \sum_{t=1}^{T} \mu(I_t) = \sum_{t=1}^{T} \Delta(I_t) \qquad (1)$$

where we denote by $\Delta(a) = \mu(a^*) - \mu(a)$ the reward gap for arm $a \in [K]$. We measure the total corruptions of rewards introduced by the adversary as

$$C = \sum_{t=1}^{T} \max_{a \in [K]} |\tilde{r}_t(a) - r_t(a)| \qquad (2)$$

which is termed as corruption level and remains unknown to the learner. Finally, we introduce the following graph-theoretical definitions (West et al. 2001).

Definition 1 (Independent Set) *An independent set $\mathcal{I}$ in a graph $G = (V, E)$ is a subset of V such that no two vertices in $\mathcal{I}$ are adjacent.*

Definition 2 (Independence Number) *The independence number α of a graph G is the cardinality of the largest independent set in G.*

Algorithm

Our algorithm belongs to the family of active arm elimination (AAE) methods and is a variant of the BARBAR algorithm (Gupta, Koren, and Talwar 2019). Before presenting our algorithm, we first review AAE and BARBAR. The basic idea of AAE is to maintain a subset $\mathcal{S} \subseteq [K]$ of arms and repeat the following two steps until T rounds of interactions are exhausted.

(a) Play each arm in $\mathcal{S}$ once and then update the empirical mean rewards for all arms in $\mathcal{S}$ based on the received feedback (which consumes $|\mathcal{S}|$ rounds of interactions in total).

(b) Eliminate arms from $\mathcal{S}$ whose empirical mean rewards are significantly worse than the maximal empirical mean reward.

If there is no corruption, it can be shown that with high probability, after $O(\ln T)$ iterations of the above two steps, the set $\mathcal{S}$ contains only optimal arms. Since each iteration involves at most $|\mathcal{S}| \leq K$ pulls of sub-optimal arms and each pull of a sub-optimal arm incurs at most 1 regret, the overall regret can be upper bounded by $O(K \ln T)$, which is minimax optimal for stochastic MAB. While AAE is effective in the purely stochastic setting, it can suffer linear regrets when adversarial corruptions of rewards exist. To see this, consider an adversary who, in some initial rounds, consistently manipulate the feedback observed by AAE in the way that the corrupted rewards of optimal arms are set to be significantly smaller than those of sub-optimal arms. After the initial rounds, optimal arms have significantly worse empirical mean rewards and are hence eliminated from $\mathcal{S}$ according to Step (b). Furthermore, Step (a) indicates that

Algorithm 1 Elise

Input: confidence $\delta \in (0,1)$, time horizon T

Initialize $m \leftarrow 1, \tau_1 \leftarrow 1, t \leftarrow 1, \tilde{\Delta}_0(a) \leftarrow 1, \forall a \in [K]$
Set $\lambda \leftarrow 273 \ln(3K\delta^{-1} \log_2 T)$
while $t \leq T$ **do**
  Invoke Algorithm 2 with inputs $\{\tilde{\Delta}_{m-1}(a)\}_{a\in[K]}$ to get an independent set $\mathcal{I}_m$
  $\tau_{m+1} \leftarrow \min\left(T+1, \tau_m + \lceil \lambda \sum_{a\in\mathcal{I}_m} \left(\tilde{\Delta}_{m-1}(a)\right)^{-2} \rceil\right)$
  Compute p_m according to (3)
  while $t < \tau_{m+1}$ **do**
    Draw an arm $I_t \sim p_m$ to play
    Observe rewards $\{\tilde{r}_t(a)\}_{a\in\phi(I_t)}$
    $t \leftarrow t+1$
  end while
  for $a = 1, \ldots, K$ **do**
    Compute an empirical mean reward $\bar{r}_m(a)$ by (4)
  end for
  Set $\hat{r}^*_m \leftarrow \max_{a\in[K]} \bar{r}_m(a) - \tilde{\Delta}_{m-1}(a)/10$
  for $a = 1, \ldots, K$ **do**
    Update the estimated reward gap as

$$\tilde{\Delta}_m(a) \leftarrow \max\left(2^{-m}, \hat{r}^*_m - \bar{r}_m(a)\right)$$

  end for
  $m \leftarrow m+1$
end while

only arms in $\mathcal{S}$ can be played. Thus, after the initial rounds, AAE will always play sub-optimal arms, which leads to linear regrets.

To make AAE robust to adversarial corruptions, Gupta, Koren, and Talwar (2019) proposed a variant algorithm of AAE called BARBAR. The main idea of BARBAR is to partition the rounds into epochs and maintain an epoch-variant estimated reward gap $\tilde{\Delta}(a) > 0$ for each arm a. In each round of an epoch, BARBAR randomly chooses an arm a from the whole arm set $[K]$ to play with probability proportional to $\left(\tilde{\Delta}(a)\right)^{-2}$. In this way, optimal arms, even with rewards heavily corrupted by the adversary, always have chances to be chosen. On the other hand, arms with significantly worse empirical mean rewards (and hence large estimated reward gaps) are unlikely to be played, which can be viewed as a soft version of arm elimination. Gupta, Koren, and Talwar (2019) proved that BARBAR attains an $O(K \ln K \ln T + KC)$ regret for stochastic MAB with adversarial corruptions, where C is the corruption level.

We here propose a variant of BARBAR for graphical bandits, which can leverage the graph-structured feedback to achieve an improved regret bound of $O(\alpha \ln K \ln T + \alpha C)$, where α is the independence number of the feedback graph G. The main difference between our algorithm and BARBAR is that in each epoch we only play arms from a carefully-designed independent set $\mathcal{I}$ of the feedback graph G instead of the whole arm set $[K]$, which reduces the cost of exploring arms once from $O(K)$ to $O(\alpha)$. While the idea of restricting exploration to an independent set has been used

Algorithm 2

Input: estimated reward gaps $\{\tilde{\Delta}(a)\}_{a\in[K]}$

Initialize $\mathcal{A} \leftarrow [K]$ and $\mathcal{I} \leftarrow \varnothing$
repeat
  Choose an arm $a \in \arg\min_{a'\in\mathcal{A}} \tilde{\Delta}(a')$
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{a\}$
  $\mathcal{A} \leftarrow \mathcal{A} - \phi(a)$
until $\mathcal{A} = \varnothing$
return $\mathcal{I}$

for stochastic graphical bandits (Cohen, Hazan, and Koren 2016), the novelty of our algorithm lies in the construction of the independent set $\mathcal{I}$. Specifically, with initializations $\mathcal{A} = [K]$ and $\mathcal{I} = \varnothing$, we construct $\mathcal{I}$ by repeating the following three steps until $\mathcal{A} = \varnothing$: choosing an arm a from $\mathcal{A}$ with the minimum estimated reward gap, adding a into $\mathcal{I}$, and removing all neighbors of a from $\mathcal{A}$. This procedure is summarized in Algorithm 2 and the intuition behind it is as follows. For each arm $a \in [K]$ that is not in $\mathcal{I}$, Algorithm 2 ensures that there must exist an arm $b \in \mathcal{I}$ with $(a,b) \in E$ and $\tilde{\Delta}(b) \leq \tilde{\Delta}(a)$. This implies that the probability of observing the reward of a (via playing b) is not smaller than the (imagined) probability of playing a. In other words, by only playing arms in $\mathcal{I}$, we can guarantee, for each arm $a \in [K]$, at least the same expected number of times of observing a's reward as that in BARBAR, which is crucial for deriving our graph-dependent logarithmic regret bound.

We now describe our algorithm in detail, which is termed as exploring on independent sets (Elise) and outlined in Algorithm 1. Let m index epoch and $\tilde{\Delta}_m(a)$ denote the estimated reward gap of arm a computed using observations during the m-th epoch. In each epoch $m = 1, 2, \ldots$, Elise first computes an independent set $\mathcal{I}_m$ by invoking Algorithm 2 with inputs $\{\tilde{\Delta}_{m-1}(a)\}_{a\in[K]}$. Then, Elise repeatedly draws an arm from $\mathcal{I}_m$ to play according to a probability distribution p_m. In the regret analysis, for each arm $a \in \mathcal{I}_m$, we require it being played around $\lambda\left(\tilde{\Delta}_{m-1}(a)\right)^{-2}$ times in expectation. To this end, we determine the epoch length to be $\left\lceil \lambda \sum_{a\in\mathcal{I}_m} \left(\tilde{\Delta}_{m-1}(a)\right)^{-2} \right\rceil$ and design the probability distribution p_m as

$$p_m(a) = \begin{cases} \frac{\left(\tilde{\Delta}_{m-1}(a)\right)^{-2}}{\sum_{a'\in\mathcal{I}_m}\left(\tilde{\Delta}_{m-1}(a')\right)^{-2}} & a \in \mathcal{I}_m \\ 0 & a \notin \mathcal{I}_m. \end{cases} \tag{3}$$

Let $O_t(a) = \mathbb{1}\{a \in \phi(I_t)\}$ be an indicator random variable representing the event that the reward of arm a is observed in round t. At the end of the m-th epoch, for each arm $a \in [K]$, Elise computes an empirical mean reward $\bar{r}_m(a)$ as

$$\bar{r}_m(a) = \frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} \tilde{r}_t(a) O_t(a)}{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)} \tag{4}$$

where we denote by τ_m the first round in the m-th epoch and use the convention $0/0 = 0$. Finally, Elise computes a lower confidence bound $\hat{r}^*_m$ for the mean reward of the optimal arm

$\mu(a^*)$ and updates the estimated reward gap $\tilde{\Delta}_m(a)$ for each arm $a \in [K]$ based on the difference between its empirical mean reward $\bar{r}_m(a)$ and the lower confidence bound $\hat{r}^*_m$.

While our algorithm Elise is an extension of BABBAR, we would like to emphasize that it also involves improvements and refinements over BARBAR as follows.

- At the end of the m-th epoch, BARBAR computes an estimated mean reward for each arm $a \in [K]$ as

$$\bar{r}'_m(a) = \frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} \tilde{r}_t(a) O_t(a)}{\mathbb{E}\big[\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)\big]}. \tag{5}$$

  BARBAR uses this estimation rather than the empirical mean reward $\bar{r}_m(a)$ defined in (4) to compute the lower confidence bound $\hat{r}^*_m$ for the mean reward of the optimal arm and update the estimated reward gaps $\big\{\tilde{\Delta}_m(a)\big\}_{a\in[K]}$. Since the denominator in (5) is the expected number of times that the reward of arm a is observed, the value of $\bar{r}'_m(a)$ can be larger than 1 and consequently the estimated reward gap $\tilde{\Delta}_m(a)$ can exceed 1. However, the regret analysis of BARBAR (proof of Lemma 4 in Gupta, Koren, and Talwar, 2019) requires $\tilde{\Delta}_m(a) \le 1$ and thus may be problematic. We address this issue by replacing the expected number of times in (5) with the actual number of times in (4), which ensures that $\big\{\bar{r}_m(a)\big\}_{a\in[K]}$ and hence $\big\{\tilde{\Delta}_m(a)\big\}_{a\in[K]}$ are upper bounded by 1.

- In the regret analyses of both BARBAR and our algorithm Elise, the core lemma is to bound the difference between the estimated (resp. empirical) mean reward $\bar{r}_m(a)$ (resp. $\bar{r}'_m(a)$) and the true mean reward $\mu(a)$ for each arm $a \in [K]$. While Gupta, Koren, and Talwar (2019) proved this core lemma by applying the multiplicative version of Chernoff-Hoeffding inequality (Dubhashi and Panconesi 2009), we derive it by employing the Bernstein inequality (Bernstein 1924). The advantage of using the Bernstein inequality is that we can configure our algorithm Elise with a smaller constant $\lambda = 273\ln(3K\delta^{-1}\log_2 T)$ compared to $\lambda = 1024\ln(8K\delta^{-1}\log_2 T)$ in BARBAR. Since in both algorithms the epoch length is proportional to λ, for the same MAB problem, our algorithm Elise has a smaller epoch length and hence a higher frequency of updating the estimated reward gaps compared to BARBAR, which also translates into better empirical performance as shown in the section of Experiments.

For Elise, we have the following theoretical guarantee.

Theorem 1 *Let* $\mathcal{A}^* = \{a \in [K] \mid \mu(a) = \max_{a'\in[K]} \mu(a')\}$ *denote the set of all optimal arms and* $\mathcal{I}^*$ *be an independent set with the maximum sum of inverse of reward gap over sub-optimal arms*

$$\mathcal{I}^* \in \arg\max_{\mathcal{I}\in \mathit{Ind}(G)} \sum_{a\in\mathcal{I}-\mathcal{A}^*} \frac{1}{\Delta(a)} \tag{6}$$

where we denote by $\mathit{Ind}(G)$ *the collection of all independent sets of the feedback graph* G*. With probability at least* $1-\delta$*, for* $T \ge 5$*, the regret of Algorithm 1 satisfies*

$$\begin{aligned}\mathrm{R}(T) &\le 1732\alpha C + 3731\ln(3K\delta^{-1}\log_2 T)\sum_{a\in\mathcal{I}^*-\mathcal{A}^*}\frac{\log_2 T}{\Delta(a)}\\ &= O(\alpha \ln K \ln T + \alpha C).\end{aligned}$$

Remark 1 Our regret bound recovers the $O(K\ln K\ln T + KC)$ regret bound of BARBAR for MAB ($\alpha = K$) and directly implies a regret bound of $O(\ln K\ln T + C)$ for PEA ($\alpha = 1$). We notice that there exists a smaller regret bound of $O(\ln K + C)$ for stochastic PEA with adversarial corruptions (Amir et al. 2020). However, this bound only holds in expectation and its proof assumes the optimal expert is unique. By contrast, we derive a high probability bound and our analysis does not require the unique assumption and applies to general scenarios with any number of optimal experts.

Remark 2 According to the theoretical analysis of BARBAR in Gupta, Koren, and Talwar (2019), the precise regret bound of BARBAR is

$$2097152\ln(8K\delta^{-1}\log_2 T)\sum_{a\in[K]-\mathcal{A}^*}\frac{\log_2 T}{\Delta(a)} + 2048KC$$

in which the leading constant factor is very large. By contrast, we optimize the constants used in the regret analysis and the leading constant factor in the regret bound of our algorithm is only 3731, reducing that of BARBAR by more than 500 times.

Remark 3 In the purely stochastic setting, the leading constant factor in the regret bound of the AAE-AlphaSample algorithm (Cohen, Hazan, and Koren 2016) is 1280. Thus, from a theoretical perspective, the cost to achieve robustness is about a factor of three.

Theoretical Analysis

In this section, we present the proof of Theorem 1.

Preliminaries

We first introduce some notations that will be used in the proof. Let M be the index of the last epoch, i.e., the epoch with $\tau_{M+1} = T+1$. We denote by $L_m = \tau_{m+1} - \tau_m$ the length of the m-th epoch. Then, for $m \in [M-1]$, we have $L_m = \left\lceil \lambda \sum_{a\in\mathcal{I}_m} \big(\tilde{\Delta}_{m-1}(a)\big)^{-2} \right\rceil$. We also define

$$\tilde{L}_m = \lambda \sum_{a\in\mathcal{I}_m} \big(\tilde{\Delta}_{m-1}(a)\big)^{-2}, \quad \omega_m = \frac{L_m}{\tilde{L}_m}. \tag{7}$$

For $a \in [K]$, we denote by $\tilde{n}_m(a) = \sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)$ the actual number of times that the reward of arm a is observed during the m-th epoch and define $n_m(a) = \lambda \sum_{a'\in\phi(a)\cap\mathcal{I}_m} \big(\tilde{\Delta}_{m-1}(a')\big)^{-2}$. For $t \in [T]$, we denote by $c_t(a) = \tilde{r}_t(a) - r_t(a)$ the corruption added to the reward of arm a in round t by the adversary. We define $C_m = \sum_{t=\tau_m}^{\tau_{m+1}-1} \max_{a\in[K]} |c_t(a)|$, which are total corruptions of rewards during the m-th epoch.

Our analysis is based on the following proposition, the proof of which is postponed to Appendix A.

Proposition 1 *The following two facts hold.*

(i) $M \leq \log_4 T$ *and* $\tilde{L}_m \geq \lambda 2^{2(m-1)}, \forall m \in [M-1]$.

(ii) With probability at least $1-\delta$*, for all arms* $a \in [K]$ *and all epochs* $m \in [M]$*, we have* $\frac{\tilde{n}_m(a)}{n_m(a)} \leq \frac{12\kappa}{11}$ *and*

$$\tilde{\Delta}_{m-1}(a) \geq \frac{8\Delta(a)}{9} - \frac{12}{5} \cdot 2^{-m} - 3\rho_{m-1}$$

where we define $\kappa = \frac{501}{500}$ *and*

$$\rho_m = \sum_{s=1}^{m} \frac{2.2C_s}{5^{m-s}\tilde{L}_s}. \tag{8}$$

Proof of Theorem 1

We are now ready to prove Theorem 1. We first decompose the regret into epochs as

$$\mathrm{R}(T) = \sum_{t=1}^{T} \Delta(I_t) = \sum_{m=1}^{M} \sum_{t=\tau_m}^{\tau_{m+1}-1} \Delta(I_t). \tag{9}$$

According to Algorithm 1, in each epoch $m \in [M]$, only arms in $\mathcal{I}_m$ can be played. Furthermore, since $\mathcal{I}_m$ is an independent set, for each arm $a \in \mathcal{I}_m$, the number of times of playing a is equivalent to the number of times that the reward of a is observed. Thus, we have for all $m \in [M]$,

$$\begin{aligned}\sum_{t=\tau_m}^{\tau_{m+1}-1} \Delta(I_t) &= \sum_{a \in \mathcal{I}_m} \Delta(a)\tilde{n}_m(a) = \sum_{a \in \mathcal{I}_m - \mathcal{A}^*} \Delta(a)\tilde{n}_m(a) \\ &\leq \frac{12\kappa}{11} \sum_{a \in \mathcal{I}_m - \mathcal{A}^*} \Delta(a) n_m(a) \end{aligned} \tag{10}$$

where the inequality is due to Proposition 1 and $\mathcal{A}^*$ is the set comprised of all optimal arms defined in Theorem 1. Fix an epoch $m \in [M]$ and a sub-optimal arm $a \in \mathcal{I}_m - \mathcal{A}^*$. There are three cases as follows.

(1) $0 < \Delta(a) \leq 4/2^m$. Then, we can bound $n_m(a)$ as

$$\begin{aligned} n_m(a) &= \lambda \sum_{a' \in \phi(a) \cap \mathcal{I}_m} \big(\tilde{\Delta}_{m-1}(a')\big)^{-2} \\ &= \lambda \big(\tilde{\Delta}_{m-1}(a)\big)^{-2} \leq \lambda 2^{2(m-1)} \leq \frac{4\lambda}{\big(\Delta(a)\big)^2} \end{aligned}$$

which leads to $\Delta(a) n_m(a) \leq \frac{4\lambda}{\Delta(a)}$.

(2) $\Delta(a) > 4/2^m$ and $\rho_{m-1} \leq \Delta(a)/36$. In this case, by Proposition 1, we have

$$\begin{aligned} \tilde{\Delta}_{m-1}(a) &\geq \frac{8\Delta(a)}{9} - \frac{12}{5} \cdot 2^{-m} - 3\rho_{m-1} \\ &\geq \left(\frac{8}{9} - \frac{3}{5} - \frac{1}{12}\right)\Delta(a) \geq \frac{\Delta(a)}{5} \end{aligned}$$

which, in turn, implies

$$\Delta(a) n_m(a) = \Delta(a)\lambda\big(\tilde{\Delta}_{m-1}(a)\big)^{-2} \leq \frac{25\lambda}{\Delta(a)}.$$

(3) $\Delta(a) > 4/2^m$ and $\rho_{m-1} > \Delta(a)/36$. Then, we write

$$\begin{aligned} \Delta(a) n_m(a) &\leq 36\rho_{m-1} n_m(a) = 36\rho_{m-1}\lambda\big(\tilde{\Delta}_{m-1}(a)\big)^{-2} \\ &\leq 36\lambda\rho_{m-1}2^{2(m-1)} \leq 9\lambda\rho_{m-1}2^{2m}. \end{aligned}$$

Combining the above three cases, we obtain that for all sub-optimal arms $a \in \mathcal{I}_m - \mathcal{A}^*$,

$$\Delta(a) n_m(a) \leq \frac{25\lambda}{\Delta(a)} + 9\lambda\rho_{m-1}2^{2m}.$$

Substituting this inequality into (10) gives

$$\begin{aligned} \sum_{t=\tau_m}^{\tau_{m+1}-1} \Delta(I_t) &\leq \frac{12\kappa}{11} \sum_{a \in \mathcal{I}_m - \mathcal{A}^*} \left(\frac{25\lambda}{\Delta(a)} + 9\lambda\rho_{m-1}2^{2m}\right) \\ &\leq \frac{300\kappa\lambda}{11} \sum_{a \in \mathcal{I}^* - \mathcal{A}^*} \frac{1}{\Delta(a)} + \frac{108\alpha\kappa\lambda}{11}\rho_{m-1}2^{2m} \end{aligned}$$

where $\mathcal{I}^*$ is the independent set with the maximum sum of inverse of reward gap over sub-optimal arms defined in (6). Combining this inequality with (9), we get

$$\begin{aligned} \mathrm{R}(T) &\leq \frac{300\kappa\lambda}{11} \sum_{a \in \mathcal{I}^* - \mathcal{A}^*} \frac{M}{\Delta(a)} + \frac{108\alpha\kappa\lambda}{11} \sum_{m=1}^{M} \rho_{m-1}2^{2m} \\ &\leq \frac{300\kappa\lambda}{22} \sum_{a \in \mathcal{I}^* - \mathcal{A}^*} \frac{\log_2 T}{\Delta(a)} + \frac{108\alpha\kappa\lambda}{11} \sum_{m=1}^{M} \rho_{m-1}2^{2m} \end{aligned}$$

where the second inequality follows from Proposition 1. It remains to bound the summation $\sum_{m=1}^{M} \rho_{m-1}2^{2m}$. By Proposition 1 and the definition of ρ_m in (8), we have

$$\begin{aligned} \sum_{m=1}^{M} \rho_{m-1}2^{2m} &= \sum_{m=1}^{M} 2^{2m} \sum_{s=1}^{m-1} \frac{2.2C_s}{5^{m-1-s}\tilde{L}_s} \\ &\leq \sum_{m=1}^{M} 2^{2m} \sum_{s=1}^{m-1} \frac{2.2C_s}{5^{m-1-s}\lambda 2^{2(s-1)}} \\ &= \frac{35.2}{\lambda} \sum_{m=1}^{M} \sum_{s=1}^{m-1} (4/5)^{m-1-s} C_s \\ &= \frac{35.2}{\lambda} \sum_{s=1}^{M-1} C_s \sum_{m=s+1}^{M} (4/5)^{m-1-s} \\ &\leq \frac{35.2}{\lambda} \sum_{s=1}^{M-1} C_s \sum_{h=0}^{+\infty} (4/5)^h \leq \frac{176C}{\lambda}. \end{aligned}$$

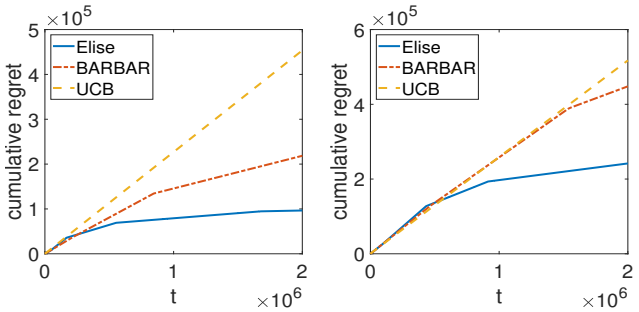
By combining the above two inequalities and recalling that $\lambda = 273\ln(3K\delta^{-1}\log_2 T)$ and $\kappa = 501/500$, we obtain

$$\mathrm{R}(T) \leq 1732\alpha C + 3731\ln(3K\delta^{-1}\log_2 T) \sum_{a \in \mathcal{I}^* - \mathcal{A}^*} \frac{\log_2 T}{\Delta(a)}.$$

which finishes the proof of Theorem 1. □

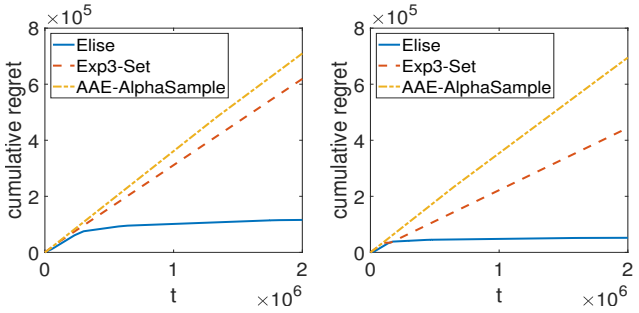
Experiments

In this section, we present numerical results to demonstrate the effectiveness of our algorithm.



(a) $K = 5$ (b) $K = 10$

Figure 1: Comparison of our algorithm versus BARBAR and UCB for MAB



(a) $p = 0.2$ (b) $p = 0.5$

Figure 2: Comparison of our algorithm versus Exp3-Set and AAE-AlphaSample for GB

Multi-Armed Bandits

To show the improvements of our algorithm Elise over BARBAR as mentioned, we first conduct experiments in the MAB setting, which is a special case of graphical bandits with $\alpha = K$. Without loss of generality, we consider an arm set in which the first two arms are optimal and the other arms are sub-optimal. For the optimal arms $a \in \{1, 2\}$, the mean reward is set as $\mu(a) = 0.8$. For sub-optimal arms $a \in \{3, \ldots, K\}$, the mean reward $\mu(a)$ is drawn uniformly from $[0.4, 0.6]$. For each arm $\in [K]$, we generate the stochastic rewards $\{r_t(a)\}_{t\in[T]}$ according to a truncated normal distribution with support $[0, 1]$, mean $\mu(a)$, and variance 0.01. We use a relatively large time horizon $T = 2000000$ and set the corruption level as $C = 1000 \ln(T)$. Following Liu and Shroff (2019), we only corrupt the reward of the optimal arm. Specifically, in each round $t \in [T]$, we set the corrupted reward of the optimal arm to be zero, if the reward of the optimal arm is to be observed by the algorithm and the total corruptions have not exceeded the corruption level C.

We compare our algorithm Elise with BARBAR as well as UCB which is designed for stochastic MAB. We run each algorithm 10 times and report the average performance in Fig. 1. As can be seen, UCB suffers the largest regret, which is expected since UCB is vulnerable to adversarial corruptions (Liu and Shroff 2019). Furthermore, Elise outperforms BARBAR in each experiment, confirming our claim that the smaller epoch length and hence the higher updating frequency can boost the performance. Finally, the performance gap between Elise and BARBAR increases with K, which is consistent with the theoretical analysis: Both algorithms enjoy an $O(K \ln K \ln T + KC)$ regret bound for the adversarially-corrupted stochastic MAB problem with K arms, but the leading constant factor in the regret bound of Elise is much smaller than that of BARBAR.

Graphical Bandits

We now turn to general graphical bandits setting with $\alpha \neq K$. We set $K = 10$ and adopt the Erdos–Renyi model to generate the feedback graph (Erdos and Renyi 1960). Specifically, for each pair of arms $(u, v) \in [K] \times [K]$ with $u \neq v$, we connect them with a fixed probability p. Intuitively, with p increasing, the feedback graph becomes denser and the independence number gets smaller. Except for the feedback graph, we follow the same experimental setup as in the above subsection.

We use two baseline algorithms in the experiment, i.e., AAE-AlphaSample (Cohen, Hazan, and Koren 2016) and Exp3-Set (Alon et al. 2013). The former is designed for stochastic graphical bandits, while the latter applies to adversarial graphical bandits. Each baseline algorithm as well as our algorithm Elise is tested 10 times and the average performance is pictured in Fig. 2. Unsurprisingly, Elise behaves the best for each configuration of p, as it can exploit the stochastic pattern of rewards and also tolerate adversarial corruptions. Furthermore, the regret of Elise decreases with p, validating the $O(\alpha \ln K \ln T + \alpha C)$ regret bound.

Conclusion and Future Work

We have formulated a new graphical bandits model termed as stochastic graphical bandits with adversarial corruptions. For this model, we proposed an online algorithm that can utilize the stochastic nature of rewards and be also robust to adversarial corruptions. Our algorithm is a non-trivial extension of the BARBAR method and involves a novel policy for constructing independent sets. Furthermore, we also made improvements and refinements of BARBAR, leading to smaller constant factors in the regret bound and better empirical performance in the experiments. Finally, as a byproduct, we provided the first high probability regret bound for stochastic PEA with adversarial corruptions.

Currently, we only consider bounded rewards. In the future, we will try to develop more robust algorithms for stochastic graphical bandits with adversarial corruptions that can handle unbounded and even heavy-tailed rewards (Bubeck, Cesa-Bianchi, and Lugosi 2013; Lu et al. 2019). Another future direction is to develop algorithms with $O(C)$ dependency in regret bounds.

Acknowledgments

This work was partially supported by the National Key R&D Program of China (2017YFB1002201), JiangsuSF (BK20200064), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. We thank the anonymous reviewers for their constructive suggestions.

References

Alon, N.; Cesa-Bianchi, N.; Dekel, O.; and Koren, T. 2015. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the 28th Conference on Learning Theory*, volume 40.

Alon, N.; Cesa-Bianchi, N.; Gentile, C.; Mannor, S.; Mansour, Y.; and Shamir, O. 2017. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing* 46(6): 1785–1826.

Alon, N.; Cesa-Bianchi, N.; Gentile, C.; and Mansour, Y. 2013. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems 26*, 1610–1618.

Altschuler, J.; Brunel, V.-E.; and Malek, A. 2019. Best arm identification for contaminated bandits. *Journal of Machine Learning Research* 20(91): 1–39.

Amir, I.; Attias, I.; Koren, T.; Livni, R.; and Mansour, Y. 2020. Prediction with Corrupted Expert Advice. *arXiv preprint arXiv:2002.10286* .

Arora, R.; Marinov, T. V.; and Mohri, M. 2019. Bandits with feedback graphs and switching costs. In *Advances in Neural Information Processing Systems 32*, 10397–10407.

Auer, P.; and Chiang, C.-K. 2016. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the 29th Conference on Learning Theory*, 116–120.

Bernstein, S. 1924. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math* 1(4): 38–49.

Bogunovic, I.; Krause, A.; and Scarlett, J. 2020. Corruption-Tolerant Gaussian Process Bandit Optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 1071–1081.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* 5(1): 1–122.

Bubeck, S.; Cesa-Bianchi, N.; and Lugosi, G. 2013. Bandits with heavy tail. *IEEE Transactions on Information Theory* 59(11): 7711–7717.

Bubeck, S.; and Slivkins, A. 2012. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Conference on Learning Theory*, 42–1.

Buccapatnam, S.; Eryilmaz, A.; and Shroff, N. B. 2014. Stochastic Bandits with Side Observations on Networks. In *Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, 289–300.

Caron, S.; Kveton, B.; Lelarge, M.; and Bhagat, S. 2012. Leveraging side observations in stochastic bandits. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 142–151.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Chen, X.; Krishnamurthy, A.; and Wang, Y. 2019. Robust Dynamic Assortment Optimization in the Presence of Outlier Customers. *arXiv preprint arXiv:1910.04183* .

Cohen, A.; Hazan, T.; and Koren, T. 2016. Online learning with feedback graphs without the graphs. In *Proceedings of the 33rd International Conference on Machine Learning*, 811–819.

Dubhashi, D. P.; and Panconesi, A. 2009. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press.

Erdos, P.; and Renyi, A. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* 5(1): 17–60.

Even-Dar, E.; Mannor, S.; and Mansour, Y. 2006. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research* 7(Jun): 1079–1105.

Feng, Z.; and Loh, P.-L. 2018. Online learning with graph-structured feedback against adaptive adversaries. In *2018 IEEE International Symposium on Information Theory (ISIT)*, 931–935.

Garcelon, E.; Roziere, B.; Meunier, L.; Teytaud, O.; Lazaric, A.; and Pirotta, M. 2020. Adversarial Attacks on Linear Contextual Bandits. *arXiv preprint arXiv:2002.03839* .

Guan, Z.; Ji, K.; Bucci Jr, D. J.; Hu, T. Y.; Palombo, J.; Liston, M.; and Liang, Y. 2020. Robust Stochastic Bandit Algorithms under Probabilistic Unbounded Adversarial Attack. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Gupta, A.; Koren, T.; and Talwar, K. 2019. Better Algorithms for Stochastic Bandits with Adversarial Corruptions. In *Proceedings of the 32nd Conference on Learning Theory*, 1562–1578.

Hu, B.; Mehta, N. A.; and Pan, J. 2019. Problem-dependent regret bounds for online learning with feedback graphs. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*.

Jun, K.-S.; Li, L.; Ma, Y.; and Zhu, J. 2018. Adversarial attacks on stochastic bandits. In *Advances in Neural Information Processing Systems 31*, 3640–3649.

Kapoor, S.; Patel, K. K.; and Kar, P. 2019. Corruption-tolerant bandit learning. *Machine Learning* 108(4): 687–715.

Kocák, T.; Neu, G.; and Valko, M. 2016. Online learning with noisy side observations. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 1186–1194.

Kocák, T.; Neu, G.; Valko, M.; and Munos, R. 2014. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems 27*, 613–621.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Lee, C.; Luo, H.; and Zhang, M. 2020. A Closer Look at Small-loss Bounds for Bandits with Graph Feedback. In

Abernethy, J. D.; and Agarwal, S., eds., *Proceedings of the 33rd Conference on Learning Theory*, 2516–2564.

Li, Y.; Lou, E. Y.; and Shan, L. 2019. Stochastic Linear Optimization with Adversarial Corruption. *arXiv preprint arXiv:1909.02109* .

Liu, F.; Buccapatnam, S.; and Shroff, N. 2018. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 3643–3650.

Liu, F.; and Shroff, N. 2019. Data Poisoning Attacks on Stochastic Bandits. In *Proceedings of the 36th International Conference on Machine Learning*, 4042–4050.

Liu, F.; Zheng, Z.; and Shroff, N. 2018. Analysis of thompson sampling for graphical bandits without the graphs. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 13–22.

Liu, G.; and Lai, L. 2020. Action-Manipulation Attacks on Stochastic Bandits. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, 3112–3116.

Lu, S.; Wang, G.; Hu, Y.; and Zhang, L. 2019. Optimal Algorithms for Lipschitz Bandits with Heavy-tailed Rewards. In *Proceedings of the 36th International Conference on Machine Learning*, 4154–4163.

Lykouris, T.; Mirrokni, V.; and Paes Leme, R. 2018. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 114–122.

Lykouris, T.; Simchowitz, M.; Slivkins, A.; and Sun, W. 2019. Corruption Robust Exploration in Episodic Reinforcement Learning. *arXiv preprint arXiv:1911.08689* .

Lykouris, T.; Sridharan, K.; and Tardos, É. 2018. Small-loss bounds for online learning with partial information. In *Proceedings of the 31st Conference on Learning Theory*, 979–986.

Lykouris, T.; Tardos, E.; and Wali, D. 2020. Feedback graph regret bounds for Thompson Sampling and UCB. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 592–614.

Ma, Y.; Jun, K.-S.; Li, L.; and Zhu, X. 2018. Data poisoning attacks in contextual bandits. In *Proceedings of the 8th International Conference on Decision and Game Theory for Security*, 186–204.

Maghsudi, S.; and Hossain, E. 2016. Multi-armed bandits with application to 5G small cells. *IEEE Wireless Communications* 23(3): 64–73.

Mannor, S.; and Shamir, O. 2011. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems 24*, 684–692.

Mourtada, J.; and Gaïffas, S. 2019. On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research* 20(83): 1–28.

Neu, G. 2015. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems 28*, 3168–3176.

Rangi, A.; and Franceschetti, M. 2019. Online learning with feedback graphs and switching costs. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2435–2444.

Schwartz, E. M.; Bradlow, E. T.; and Fader, P. S. 2017. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science* 36(4): 500–522.

Seldin, Y.; and Lugosi, G. 2017. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 30th Conference on Learning Theory*, 1743–1759.

Seldin, Y.; and Slivkins, A. 2014. One Practical Algorithm for Both Stochastic and Adversarial Bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 1287–1295.

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

Tossou, A. C.; Dimitrakakis, C.; and Dubhashi, D. 2017. Thompson sampling for stochastic bandits with graph feedback. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

Villar, S. S.; Bowden, J.; and Wason, J. 2015. Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science* 30(2): 199–215.

Wei, C.-Y.; and Luo, H. 2018. More Adaptive Algorithms for Adversarial Bandits. In *Proceedings of the 31st Conference On Learning Theory*, 1263–1291.

West, D. B.; et al. 2001. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River.

Wilbur, K. C.; and Zhu, Y. 2009. Click fraud. *Marketing Science* 28(2): 293–308.

Zimmert, J.; Luo, H.; and Wei, C.-Y. 2019. Beating Stochastic and Adversarial Semi-bandits Optimally and Simultaneously. In *Proceedings of the 36th International Conference on Machine Learning*, 7683–7692.

Zimmert, J.; and Seldin, Y. 2019. An Optimal Algorithm for Stochastic and Adversarial Bandits. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 467–475.

A Proof of Proposition 1

Following Gupta, Koren, and Talwar (2019), we first analyze the length L_m of each epoch m. To this end, recalling that $L_m = \omega_m \tilde{L}_m$, we establish bounds for $\tilde{L}_m$ and ω_m as follows.

Lemma 1 *For all epochs $m \in [M-1]$, we have*

$$\tilde{L}_m \geq \lambda 2^{2(m-1)}, \qquad 1 \leq \omega_m \leq \kappa \tag{11}$$

where κ is defined in Proposition 1. Furthermore, the number of epochs satisfies $M \leq \log_4 T$.

Proof. For $m \geq 2$, according to Step 15 of Algorithm 1, there exists an arm $a^*_{m-1} \in [K]$ satisfying $\hat{r}^*_{m-1} = \bar{r}_{m-1}(a^*_{m-1}) - \tilde{\Delta}_{m-2}(a^*_{m-1})/10$, which, together with Step 17 of Algorithm 1, implies

$$\tilde{\Delta}_{m-1}(a^*_{m-1}) = \max\left(2^{-(m-1)}, \hat{r}^*_{m-1} - \bar{r}_{m-1}(a^*_{m-1})\right) \leq \max\left(2^{-(m-1)}, 0\right) = 2^{-(m-1)}.$$

On the other hand, for each arm $a \in [K]$, Step 17 of Algorithm 1 implies $\tilde{\Delta}_{m-1}(a) \geq 2^{-(m-1)}$. Thus, according to Step 4 of Algorithm 1 and the construction of $\mathcal{I}_m$ in Algorithm 2, there must exist an arm $a_m \in \mathcal{I}_m$ with $\tilde{\Delta}_{m-1}(a_m) = 2^{-(m-1)}$. Note that this property also holds for $m = 1$, as $\tilde{\Delta}_0(a)$ is initialized to be 1 for all arms $a \in [K]$. So, we have for each $m \in [M-1]$,

$$\tilde{L}_m \overset{(7)}{=} \lambda \sum_{a \in \mathcal{I}_m} \left(\tilde{\Delta}_{m-1}(a)\right)^{-2} \geq \lambda \left(\tilde{\Delta}_{m-1}(a_m)\right)^{-2} = \lambda 2^{2(m-1)}.$$

The above inequality also implies

$$\omega_m = \frac{L_m}{\tilde{L}_m} = \frac{\lceil \tilde{L}_m \rceil}{\tilde{L}_m} \leq 1 + \frac{1}{\lambda} = 1 + \frac{1}{273 \ln(3K\delta^{-1} \log_2 T)} \leq 1 + \frac{1}{500} = \kappa$$

where the last inequality is due to $T \geq 5$. The lower bound of ω_m is obvious as $L_m = \lceil \tilde{L}_m \rceil \geq \tilde{L}_m$. It remains to bound M. Since for each epoch $m \in [M-1]$, its length satisfies $L_m \geq \tilde{L}_m \geq \lambda 2^{2(m-1)}$, we have

$$T \geq \sum_{m=1}^{M-1} L_m \geq \sum_{m=1}^{M-1} \lambda 2^{2(m-1)}.$$

By solving this inequality with $T \geq 5$ and $\lambda \geq 500$, we obtain $M \leq \log_4 T$. □

Then, we turn to bound the difference between the empirical mean reward $\bar{r}_m(a)$ and the true mean $\mu(a)$ for all arms $a \in [K]$ and all epochs $m \in [M-1]$. We introduce the following three lemmas, from which we can obtain a high probability bound on $|\bar{r}_m - \mu(a)|$. In the proofs of these lemmas, we denote by $\mathcal{H}_t$ the history observed by the learner up to and including round t and by $\mathcal{F}_t$ the history observed by the adversary up to the time of determining the corrupted rewards of round $t+1$. Formally, $\mathcal{H}_t$ and $\mathcal{F}_t$ are sigma-fields generated by the random variables $\{I_s\}_{s\in[t]}, \{\tilde{r}_s(a)\}_{s\in[t],a\in\phi(I_s)}$ and $\{I_s\}_{s\in[t]}, \{\tilde{r}_s(a)\}_{s\in[t],a\in[K]}, \{r_s(a)\}_{s\in[t+1],a\in[K]}$, respectively.

Lemma 2 *With probability at least $1 - \frac{\delta}{3}$, for all arms $a \in [K]$, the ratio between $\tilde{n}_m(a)$ and $n_m(a)$ satisfies*

$$\frac{10}{11} \leq \frac{\tilde{n}_m(a)}{n_m(a)} \leq \frac{12\kappa}{11}, \ \forall m \in [M-1] \qquad \textit{and} \qquad \frac{\tilde{n}_M(a)}{n_M(a)} \leq \frac{12\kappa}{11}. \tag{12}$$

Proof. Fix an arm $a \in [K]$ and an epoch $m \in [M-1]$. Conditioned on $\mathcal{H}_{\tau_m - 1}$, the values of both τ_m and τ_{m+1} are deterministic and the random variables $\{O_t(a)\}_{t=\tau_m,\ldots,\tau_{m+1}-1}$ are independent and follow the same Bernoulli distribution with parameter $\sum_{a' \in \phi(a) \cap \mathcal{I}_m} p_m(a')$, where p_m is defined in (3). For each $t = \tau_m, \ldots, \tau_{m+1} - 1$, we define a random variable

$$X_t = O_t(a) - \sum_{a' \in \phi(a) \cap \mathcal{I}_m} p_m(a').$$

Then, we have $\mathbb{E}[X_t \mid \mathcal{H}_{\tau_m - 1}] = 0$, $|X_t| \leq 1$, and

$$\begin{aligned}
\sum_{t=\tau_m}^{\tau_{m+1}-1} \mathbb{E}\left[X_t^2 \mid \mathcal{H}_{\tau_m - 1}\right] &\leq \sum_{t=\tau_m}^{\tau_{m+1}-1} \sum_{a' \in \phi(a) \cap \mathcal{I}_m} p_m(a') \\
&= L_m \sum_{a' \in \phi(a) \cap \mathcal{I}_m} p_m(a') = \omega_m \tilde{L}_m \sum_{a' \in \phi(a) \cap \mathcal{I}_m} p_m(a') \\
&\overset{(3,7)}{=} \omega_m \lambda \sum_{b \in \mathcal{I}_m} \left(\tilde{\Delta}_{m-1}(b)\right)^{-2} \sum_{a' \in \phi(a) \cap \mathcal{I}_m} \frac{\left(\tilde{\Delta}_{m-1}(a')\right)^{-2}}{\sum_{a'' \in \mathcal{I}_m} \left(\tilde{\Delta}_{m-1}(a'')\right)^{-2}} \\
&= \omega_m \lambda \sum_{a' \in \phi(a) \cap \mathcal{I}_m} \left(\tilde{\Delta}_{m-1}(a')\right)^{-2} = \omega_m n_m(a).
\end{aligned} \tag{13}$$

Thus, we can apply the Bernstein inequality (Bernstein 1924) and obtain that for any $\epsilon > 0$,

$$\Pr\left\{\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} X_t\right| \geq \epsilon \;\middle|\; \mathcal{H}_{\tau_m-1}\right\} \leq 2\exp\left(-\frac{\epsilon^2/2}{\omega_m n_m(a) + \epsilon/3}\right).$$

Picking $\epsilon = \frac{\omega_m n_m(a)}{11}$ leads to

$$\begin{aligned}\Pr\left\{\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} X_t\right| \geq \frac{\omega_m n_m(a)}{11} \;\middle|\; \mathcal{H}_{\tau_m-1}\right\} &\leq 2\exp\left(-\frac{\big(\omega_m n_m(a)\big)^2/242}{\omega_m n_m(a) + \omega_m n_m(a)/33}\right) \\ &\leq 2\exp(-n_m(a)/250).\end{aligned}$$

On the other hand, by the definition of X_t, we get

$$\sum_{t=\tau_m}^{\tau_{m+1}-1} X_t = \sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a) - \sum_{t=\tau_m}^{\tau_{m+1}-1} \sum_{a' \in \phi(a) \cap \mathcal{I}_m} p_m(a') \overset{(13)}{=} \tilde{n}_m(a) - \omega_m n_m(a).$$

Combining the above two inequalities gives

$$\Pr\left\{|\tilde{n}_m(a) - \omega_m n_m(a)| \geq \omega_m n_m(a)/11 \;\middle|\; \mathcal{H}_{\tau_m-1}\right\} \leq 2\exp(-n_m(a)/250).$$

By Steps 13-17 of Algorithm 1, for each arm $a' \in [K]$, its estimated reward gap satisfies $\tilde{\Delta}_{m-1}(a') \leq 1$. On the other hand, for the fixed arm a, according to Step 4 of Algorithm 1 and the construction of $\mathcal{I}_m$ in Algorithm 2, there must exist an arm $b \in \phi(a) \cap \mathcal{I}_m$. So, we have

$$n_m(a) = \lambda \sum_{a' \in \phi(a) \cap \mathcal{I}_m} \big(\tilde{\Delta}_{m-1}(a')\big)^{-2} \geq \lambda\big(\tilde{\Delta}_{m-1}(b)\big)^{-2} \geq \lambda \tag{14}$$

and hence

$$\Pr\left\{|\tilde{n}_m(a) - \omega_m n_m(a)| \geq \omega_m n_m(a)/11 \mid \mathcal{H}_{\tau_m-1}\right\} \leq 2\exp(-\lambda/250).$$

By rearranging this inequality and removing the condition, we get

$$\Pr\left\{\left|\frac{\tilde{n}_m(a)}{\omega_m n_m(a)} - 1\right| \geq \frac{1}{11}\right\} \leq 2\exp(-\lambda/250).$$

The above analysis is for epochs $m \in [M-1]$. Now, we consider the last epoch $m = M$. Let $L' = \lambda \sum_{a \in \mathcal{I}_M} \big(\tilde{\Delta}_{M-1}(a)\big)^{-2}$ and $\omega' = \frac{\lceil L' \rceil}{L'}$. We introduce a sequence of auxiliary random variables $\{O_t(a)\}_{t=T+1,\ldots,\tau_M+\lceil L' \rceil-1}$, which are i.i.d. and follow the Bernoulli distribution with parameter $\sum_{a' \in \phi(a) \cap \mathcal{I}_M} p_M(a')$. We also introduce an upper bound of $\tilde{n}_M(a)$ as

$$\hat{n}_M(a) = \sum_{t=\tau_M}^{\tau_M+\lceil L' \rceil-1} O_t(a) \geq \sum_{t=\tau_M}^{T} O_t(a) = \tilde{n}_M(a).$$

Following the same derivation as we did for epochs $m \in [M-1]$, we can obtain

$$\Pr\left\{\left|\frac{\hat{n}_M(a)}{\omega' n_M(a)} - 1\right| \geq \frac{1}{11}\right\} \leq 2\exp(-\lambda/250).$$

By substituting $\lambda = 273\ln(3K\delta^{-1}\log_2 T)$ into the above two inequalities, taking a union bound over K arms and $\lfloor \log_4 T \rfloor$ epochs, and using $\hat{n}_M(a) \geq \tilde{n}_M(a)$ and $\omega' \leq \kappa$. we finish the proof. □

Lemma 3 *With probability at least $1 - \frac{\delta}{3}$, for all arms $a \in [K]$ and epochs $m \in [M-1]$, we have*

$$\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)\big(r_t(a) - \mu(a)\big)\right| \leq \frac{2\sqrt{\lambda n_m(a)}}{23}.$$

Proof. Fix an arm $a \in [K]$ and an epoch $m \in [M-1]$. For each $t = \tau_m, \ldots, \tau_{m+1} - 1$, we define a random variable

$$X_t = O_t(a)\big(r_t(a) - \mu(a)\big).$$

Conditioned on $\mathcal{H}_{\tau_m-1}$, the values of both τ_m and τ_{m+1} are deterministic and the random variables $\{X_t\}_{t=\tau_m,\dots,\tau_{m+1}-1}$ are independent. Furthermore, we have $\mathbb{E}[X_t \mid \mathcal{H}_{\tau_m-1}] = 0$, $|X_t| \le 1$, and

$$\sum_{t=\tau_m}^{\tau_{m+1}-1} \mathbb{E}\left[X_t^2 \mid \mathcal{H}_{\tau_m-1}\right] \le \sum_{t=\tau_m}^{\tau_{m+1}-1} \mathbb{E}\left[\big(O_t(a)\big)^2 \mid \mathcal{H}_{\tau_m-1}\right] = L_m \sum_{a'\in\phi(a)\cap\mathcal{I}_m} p_m(a') \overset{(13)}{=} \omega_m n_m(a).$$

Thus, we can apply the Bernstein inequality (Bernstein 1924) and obtain that for any $\epsilon > 0$,

$$\Pr\left\{\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} X_t\right| \ge \epsilon \;\middle|\; \mathcal{H}_{\tau_m-1}\right\} \le 2\exp\left(-\frac{\epsilon^2/2}{\omega_m n_m(a) + \epsilon/3}\right).$$

Picking $\epsilon = \frac{2\sqrt{\lambda n_m(a)}}{23}$ leads to

$$\begin{aligned}\Pr\left\{\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} X_t\right| \ge \frac{2\sqrt{\lambda n_m(a)}}{23} \;\middle|\; \mathcal{H}_{\tau_m-1}\right\} &\le 2\exp\left(-\frac{2\lambda n_m(a)/529}{\omega_m n_m(a) + 2\sqrt{\lambda n_m(a)}/69}\right)\\ &\overset{(14)}{\le} 2\exp\left(-\frac{2\lambda n_m(a)/529}{\omega_m n_m(a) + 2n_m(a)/69}\right)\\ &\le 2\exp\left(-\frac{2\lambda/529}{\kappa + 2/69}\right) \le 2\exp(-\lambda/273).\end{aligned}$$

Finally, we substitute $X_t = O_t(a)\big(r_t(a) - \mu(a)\big)$ and $\lambda = 273\ln(3K\delta^{-1}\log_2 T)$ into the above inequality, remove the condition, and take a union bound over K arms and $\lfloor \log_4 T \rfloor$ epochs to complete the proof. □

Lemma 4 *With probability at least $1 - \frac{\delta}{3}$, for all arms $a \in [K]$ and epochs $m \in [M-1]$, we have*

$$\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)c_t(a)\right| \le \frac{2n_m(a)C_m}{\tilde{L}_m} + \frac{\lambda}{253}.$$

Proof. Fix an arm $a \in [K]$. For $t \in [T]$, let $\pi(t)$ denote the index of the epoch that contains round t, i.e., $\tau_{\pi(t)} \le t < \tau_{\pi(t)+1}$. We define

$$q_t(a) = \sum_{a'\in\phi(a)\cap\mathcal{I}_{\pi(t)}} p_{\pi(t)}(a'), \quad X_t = \big(O_t(a) - q_t(a)\big)c_t(a).$$

Conditioned on $\mathcal{F}_{t-1}$, both $q_t(a)$ and $c_t(a)$ are deterministic. So, we have

$$\mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] = \big(\mathbb{E}\left[O_t(a) \mid \mathcal{F}_{t-1}\right] - q_t(a)\big) \cdot c_t(a) = \big(q_t(a) - q_t(a)\big) \cdot c_t(a) = 0 \tag{15}$$

which implies that $\{X_t\}_{t\in[T]}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t\in[T]}$. Fix an epoch $m \in [M-1]$. We have $|X_t| \le |c_t(a)| \le 1$ and

$$\sum_{t=\tau_m}^{\tau_{m+1}-1} \mathbb{E}\big[X_t^2 \mid \mathcal{F}_{t-1}\big] \le \sum_{t=\tau_m}^{\tau_{m+1}-1} \big(c_t(a)\big)^2 q_t(a) \le \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)| q_t(a).$$

Thus, we can apply the Freedman-type concentration inequality for martingales (Theorem 10 of Gupta, Koren, and Talwar (2019)) and obtain that for any $\epsilon \in (0,1)$,

$$\begin{aligned}\epsilon &\ge \Pr\left\{\sum_{t=\tau_m}^{\tau_{m+1}-1} X_t \ge \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)|q_t(a) + \ln(1/\epsilon)\right\}\\ &= \Pr\left\{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)c_t(a) - \sum_{t=\tau_m}^{\tau_{m+1}-1} q_t(a)c_t(a) \ge \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)|q_t(a) + \ln(1/\epsilon)\right\}\\ &\ge \Pr\left\{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)c_t(a) \ge 2\sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)|q_t(a) + \ln(1/\epsilon)\right\}.\end{aligned}$$

Similarly, we can get

$$\epsilon \ge \Pr\left\{-\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)c_t(a) \ge 2\sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)|q_t(a) + \ln(1/\epsilon)\right\}$$

by defining $X_t = \big(q_t(a) - O_t(a)\big)c_t(a)$ instead. Thus, we have

$$\Pr\left\{\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)c_t(a)\right| \geq 2\sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)|q_t(a) + \ln(1/\epsilon)\right\} \leq 2\epsilon.$$

By picking $\epsilon = \exp(-\lambda/253)$ and noticing

$$\begin{aligned}
\sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)|q_t(a) &= \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)| \sum_{a'\in\phi(a)\cap\mathcal{I}_{\pi(t)}} p_{\pi(t)}(a') = \sum_{a'\in\phi(a)\cap\mathcal{I}_m} p_m(a') \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)| \\
&\overset{(3)}{=} \sum_{a'\in\phi(a)\cap\mathcal{I}_m} \frac{\big(\tilde{\Delta}_{m-1}(a')\big)^{-2}}{\sum_{a''\in\mathcal{I}_m}\big(\tilde{\Delta}_{m-1}(a'')\big)^{-2}} \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)| \\
&= \frac{\lambda\sum_{a'\in\phi(a)\cap\mathcal{I}_m}\big(\tilde{\Delta}_{m-1}(a')\big)^{-2}}{\lambda\sum_{a''\in\mathcal{I}_m}\big(\tilde{\Delta}_{m-1}(a'')\big)^{-2}} \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)| \\
&\overset{(7)}{=} \frac{n_m(a)}{\tilde{L}_m} \sum_{t=\tau_m}^{\tau_{m+1}-1} |c_t(a)| \leq \frac{n_m(a)C_m}{\tilde{L}_m},
\end{aligned}$$

we get

$$\Pr\left\{\left|\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)c_t(a)\right| \geq \frac{2n_m(a)C_m}{\tilde{L}_m} + \frac{\lambda}{253}\right\} \leq 2\exp(-\lambda/253). \tag{16}$$

We finish the proof by substituting $\lambda = 273\ln(3K\delta^{-1}\log_2 T)$ into the above inequality and taking a union bound over K arms and $\lfloor\log_4 T\rfloor$ epochs. □

Armed with Lemmas 2, 3, and 4, we are now ready to prove the following core lemma.

Lemma 5 *Let $\mathcal{E}$ be an event defined as*

$$\begin{aligned}
\mathcal{E} = \Bigg\{&\forall m \in [M-1], \forall a \in [K],\ \frac{10}{11} \leq \frac{\tilde{n}_m(a)}{n_m(a)} \leq \frac{12\kappa}{11} \quad \textit{and} \quad \forall a \in [K],\ \frac{\tilde{n}_M(a)}{n_M(a)} \leq \frac{12\kappa}{11} \\
&\textit{and} \quad \forall m \in [M-1], \forall a \in [K],\ |\bar{r}_m(a) - \mu(a)| \leq \frac{\tilde{\Delta}_{m-1}(a)}{10} + \frac{2.2C_m}{\tilde{L}_m}\Bigg\}.
\end{aligned} \tag{17}$$

Then, $\mathcal{E}$ holds with probability at least $1-\delta$.

Proof. We define some auxiliary events as follows.

$$\begin{aligned}
\mathcal{E}_1 &= \left\{\forall m \in [M-1], \forall a \in [K],\ \frac{10}{11} \leq \frac{\tilde{n}_m(a)}{n_m(a)} \leq \frac{12\kappa}{11} \quad \text{and} \quad \forall a \in [K],\ \frac{\tilde{n}_M(a)}{n_M(a)} \leq \frac{12\kappa}{11}\right\} \\
\mathcal{E}_2 &= \left\{\forall m \in [M-1], \forall a \in [K],\ \left|\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)\big(r_t(a) - \mu(a)\big)\right| \leq \frac{2\sqrt{\lambda n_m(a)}}{23}\right\} \\
\mathcal{E}_3 &= \left\{\forall m \in [M-1], \forall a \in [K],\ \left|\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)c_t(a)\right| \leq \frac{2n_m(a)C_m}{\tilde{L}_m} + \frac{\lambda}{253}\right\}
\end{aligned}$$

By taking a union bound over Lemmas 2, 3, and 4, we have

$$\Pr(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - \delta.$$

Below, we prove that $\Pr(\mathcal{E}) \geq \Pr(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)$. Suppose the events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ occur simultaneously. We show that the

event $\mathcal{E}$ also occurs. Specifically, we first write

$$
\begin{aligned}
|\bar{r}_m(a)-\mu(a)| &= \left|\frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)\tilde{r}_t(a)}{\tilde{n}_m(a)}-\mu(a)\right| \\
&= \left|\frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a) r_t(a)}{\tilde{n}_m(a)}+\frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a) c_t(a)}{\tilde{n}_m(a)}-\mu(a)\right| \\
&\le \left|\frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)\big(r_t(a)-\mu(a)\big)}{\tilde{n}_m(a)}\right|+\left|\frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a) c_t(a)}{\tilde{n}_m(a)}\right| \\
&= \frac{n_m(a)}{\tilde{n}_m(a)}\cdot\left\{\left|\frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a)\big(r_t(a)-\mu(a)\big)}{n_m(a)}\right|+\left|\frac{\sum_{t=\tau_m}^{\tau_{m+1}-1} O_t(a) c_t(a)}{n_m(a)}\right|\right\} \\
&\le \frac{11}{10}\cdot\frac{2}{23}\cdot\sqrt{\frac{\lambda}{n_m(a)}}+\frac{11}{10}\cdot\frac{2C_m}{\tilde{L}_m}+\frac{11}{10}\cdot\frac{1}{253}\cdot\frac{\lambda}{n_m(a)} \\
&\overset{(14)}{\le} \frac{11}{10}\cdot\frac{2}{23}\cdot\sqrt{\frac{\lambda}{n_m(a)}}+\frac{11}{10}\cdot\frac{2C_m}{\tilde{L}_m}+\frac{11}{10}\cdot\frac{1}{253}\cdot\sqrt{\frac{\lambda}{n_m(a)}} \\
&= \frac{1}{10}\cdot\sqrt{\frac{\lambda}{n_m(a)}}+\frac{2.2C_m}{\tilde{L}_m}.
\end{aligned}
\tag{18}
$$

Then, we bound $n_m(a)$ in terms of λ and $\tilde{\Delta}_{m-1}(a)$. According to Step 4 of Algorithm 1 and the construction of $\mathcal{I}_m$ in Algorithm 2, there must exist an arm $b\in\phi(a)\cap\mathcal{I}_m$ with $\tilde{\Delta}_{m-1}(b)\le\tilde{\Delta}_{m-1}(a)$, which leads to

$$
n_m(a)=\lambda\sum_{a'\in\phi(a)\cap\mathcal{I}_m}\big(\tilde{\Delta}_{m-1}(a')\big)^{-2}\ge\lambda\big(\tilde{\Delta}_{m-1}(b)\big)^{-2}\ge\lambda\big(\tilde{\Delta}_{m-1}(a)\big)^{-2}.
$$

Combining this inequality with (18), we finish the proof. □

Equipped with Lemma 5, we can analyze the lower confidence bound $\hat{r}_m^*$ for the mean reward of the optimal arm $\mu(a^*)$ computed in Step 15 of Algorithm 1.

Lemma 6 *Suppose the event $\mathcal{E}$ defined in Lemma 5 occurs. For all epochs $m\in[M-1]$, we have*

$$
-\frac{\tilde{\Delta}_{m-1}(a^*)}{5}-\frac{2.2C_m}{\tilde{L}_m}\le\hat{r}_m^*-\mu(a^*)\le\frac{2.2C_m}{\tilde{L}_m}.
$$

Proof. According to Step 15 of Algorithm 1, the lower confidence bound $\hat{r}_m^*$ satisfies

$$
\hat{r}_m^*\ge\bar{r}_m(a^*)-\frac{\tilde{\Delta}_{m-1}(a^*)}{10}\overset{(17)}{\ge}\mu(a^*)-\frac{\tilde{\Delta}_{m-1}(a^*)}{5}-\frac{2.2C_m}{\tilde{L}_m}.
$$

On the other hand, we have

$$
\begin{aligned}
\hat{r}_m^*=\max_{a\in[K]}\ \bar{r}_m(a)-\frac{\tilde{\Delta}_{m-1}(a)}{10} &\overset{(17)}{\le} \max_{a\in[K]}\ \mu(a)+\frac{\tilde{\Delta}_{m-1}(a)}{10}+\frac{2.2C_m}{\tilde{L}_m}-\frac{\tilde{\Delta}_{m-1}(a)}{10} \\
&\le \max_{a\in[K]}\ \mu(a)+\frac{2.2C_m}{\tilde{L}_m}=\mu(a^*)+\frac{2.2C_m}{\tilde{L}_m}.
\end{aligned}
$$

Combining the above two inequalities, we finish the proof. □

We now present our last lemma which bounds the estimated reward gap $\tilde{\Delta}_m(a)$ from below and above.

Lemma 7 *Suppose the event $\mathcal{E}$ defined in Lemma 5 occurs. For all arms $a\in[K]$ and epochs $m\in[M-1]$, we have*

$$
\frac{8\Delta(a)}{9}-\frac{6}{5}\cdot 2^{-m}-3\rho_m\le\tilde{\Delta}_m(a)\le\frac{10\Delta(a)}{9}+2^{-m+1}+2\rho_m
$$

where ρ_m is defined in Proposition 1. Furthermore, the above inequalities trivially holds for $m=0$, since $\Delta_0(a)=1,\forall a\in[K]$ and $\rho_0=0$.

Proof. We first prove the upper bound by induction. For $m = 1$, the upper bound follows from

$$\tilde{\Delta}_m(a) \leq 1 = 2^{-1+1}, \forall a \in [K].$$

Suppose the upper bound holds for $m - 1 \geq 1$. By Lemmas 5 and 6, we have

$$\begin{aligned}
\hat{r}^*_m - \bar{r}_m(a) &= \hat{r}^*_m - \mu(a^*) + \mu(a^*) - \mu(a) + \mu(a) - \bar{r}_m(a) \\
&\leq \frac{2.2C_m}{\tilde{L}_m} + \Delta(a) + \frac{\tilde{\Delta}_{m-1}(a)}{10} + \frac{2.2C_m}{\tilde{L}_m} \\
&\leq \Delta(a) + \frac{4.4C_m}{\tilde{L}_m} + \frac{1}{10}\left(\frac{10\Delta(a)}{9} + 2^{-(m-1)+1} + 2\rho_{m-1}\right) \\
&= \frac{10\Delta(a)}{9} + \frac{2^{-(m-1)+1}}{10} + \frac{4.4C_m}{\tilde{L}_m} + \frac{2}{10} \cdot \sum_{s=1}^{m-1} \frac{2.2C_s}{5^{m-1-s}\tilde{L}_s} \\
&\leq \frac{10\Delta(a)}{9} + 2^{-(m-1)+1} \cdot 2^{-1} + \frac{4.4C_m}{\tilde{L}_m} + \sum_{s=1}^{m-1} \frac{4.4C_s}{5^{m-s}\tilde{L}_s} \\
&= \frac{10\Delta(a)}{9} + 2^{-m+1} + 2\rho_m.
\end{aligned}$$

We finish the proof of the upper bound by recalling that

$$\tilde{\Delta}_m(a) = \max\left(2^{-m}, \hat{r}^*_m - \bar{r}_m(a)\right).$$

Then, we turn to the lower bound. By Lemmas 5 and 6 and the upper bound we have just proved, we get

$$\begin{aligned}
\hat{r}^*_m - \bar{r}_m(a) &\geq \mu(a^*) - \frac{\tilde{\Delta}_{m-1}(a^*)}{5} - \frac{2.2C_m}{\tilde{L}_m} - \left(\mu(a) + \frac{\tilde{\Delta}_{m-1}(a)}{10} + \frac{2.2C_m}{\tilde{L}_m}\right) \\
&= \Delta(a) - \frac{4.4C_m}{\tilde{L}_m} - \left(\frac{\tilde{\Delta}_{m-1}(a^*)}{5} + \frac{\tilde{\Delta}_{m-1}(a)}{10}\right) \\
&\geq \Delta(a) - \frac{4.4C_m}{\tilde{L}_m} - \left(\frac{2^{-m+2} + 2\rho_{m-1}}{5} + \frac{\Delta(a)}{9} + \frac{2^{-m+2} + 2\rho_{m-1}}{10}\right) \\
&\geq \frac{8\Delta(a)}{9} - \frac{6}{5} \cdot 2^{-m} - 3\rho_m.
\end{aligned} \tag{19}$$

As $\tilde{\Delta}_m(a) = \max\left(2^{-m}, \hat{r}^*_m - \bar{r}_m(a)\right) \geq \hat{r}^*_m - \bar{r}_m(a)$, we finish the proof of the lower bound. □

Finally, we point out that Proposition 1 directly follows from Lemmas 1, 2, and 7.