Adversarial Linear Contextual Bandits with Graph-Structured Side Observations

Lingda Wang[1], **Bingcong Li**[2], **Huozhi Zhou**[1], **Georgios B. Giannakis**[2],
Lav R. Varshney[1], **Zhizhen Zhao**[1]
[1]University of Illinois at Urbana-Champaign
[2]University of Minnesota - Twin Cities
{lingdaw2, hzhou35, varshney, zhizhenz}@illinois.edu, {lixx5599, georgios}@umn.edu

Abstract

This paper studies the adversarial graphical contextual bandits, a variant of adversarial multi-armed bandits that leverage two categories of the most common side information: *contexts* and *side observations*. In this setting, a learning agent repeatedly chooses from a set of K actions after being presented with a d-dimensional context vector. The agent not only incurs and observes the loss of the chosen action, but also observes the losses of its neighboring actions in the observation structures, which are encoded as a series of feedback graphs. This setting models a variety of applications in social networks, where both contexts and graph-structured side observations are available. Two efficient algorithms are developed based on `EXP3`. Under mild conditions, our analysis shows that for undirected feedback graphs the first algorithm, `EXP3-LGC-U`, achieves a sub-linear regret with respect to the time horizon and the average *independence number* of the feedback graphs. A slightly weaker result is presented for the directed graph setting as well. The second algorithm, `EXP3-LGC-IX`, is developed for a special class of problems, for which the regret is the same for both directed as well as undirected feedback graphs. Numerical tests corroborate the efficiency of proposed algorithms.

1 Introduction

Multi-armed bandits (MAB) (Thompson 1933; Lai and Robbins 1985; Auer, Cesa-Bianchi, and Fischer 2002; Auer et al. 2002) is an online learning model of paramount importance for sequential decision making. Yielding algorithms with both theoretical guarantees and convenient implementations such as `UCB1` (Auer, Cesa-Bianchi, and Fischer 2002), Thompson sampling (Agrawal and Goyal 2012; Kaufmann, Korda, and Munos 2012; Thompson 1933), `EXP3` (Auer et al. 2002), and `INF` (Audibert and Bubeck 2009), MAB has been of interest in many real-world applications: clinical trials (Thompson 1933), web advertising (Jiang 2015; Li, Chen, and Giannakis 2019; Zhou et al. 2020), web search (Kveton et al. 2015; Wang et al. 2019), and cognitive radio (Maghsudi and Hossain 2016), to just name a few. While the classical MAB has received much attention, this model may not be delicate enough for applications, since it does not fully leverage the widely available side information. This has motivated studies on *contextual bandits* (Li et al. 2010; Chu et al. 2011; Abbasi-Yadkori, Pál, and Szepesvári 2011; Agarwal et al. 2014) and *graphical bandits* (Mannor and Shamir 2011; Alon et al. 2013, 2015, 2017), which aim to address two categories of the most common side information, *contexts* and *side observations*, respectively. In a contextual bandit problem, a learning agent chooses an action to play based on the context for the current time step and the past interactions. In a graphical bandit setup, playing an action not only discloses its own loss, but also the losses of its neighboring actions. Applications of contextual bandits include mobile health (Tewari and Murphy 2017) and online personalized recommendation (Li et al. 2010), whereas applications of graphical bandits include viral marketing, online pricing, and online recommendation in social networks (Alon et al. 2017; Liu, Zheng, and Shroff 2018).

However, contextual or graphical bandits alone may still not capture many aspects of real-world applications in social networks efficiently. As a motivating example, consider the viral marketing over a social network (Lobel, Sadler, and Varshney 2017), where a salesperson aims to investigate the popularity of a series of products. At each time step, the salesperson could offer a survey (context) of some product to a user together with a promotion. The salesperson also has a chance to survey the user's followers (side observations) in this social network, which can be realized by assuming that i) if the user would like to get the promotion, the user should finish the questionnaire and share it in the social network, and ii) if the followers would like to get the same promotion, they need to finish the same questionnaire shared by the user.

This example demonstrates the importance of a new MAB model that accounts for both context and side observations. Thus, designing pertinently efficient algorithms with guarantees is valuable, which is also recognized in the recent work on stochastic graphical contextual bandits (Singh et al. 2020). Mechanically combining existing algorithms for contextual bandits and graphical bandits leads to algorithms with better empirical performance, compared to algorithms designed solely for contextual or graphical bandits. This can be verified by the genie-aided argument that side observations provide more information beyond the original contextual bandit problem, and will therefore not result in worse performance, if used properly. Certain theoretical guarantees can be derived if we adopt the results of contextual bandits as the worst case in the analysis. However, one should keep in mind that the merits of this paper are not just in combining formulations and algorithms: we will show that simply combining existing algorithms will result in intractable steps in analysis, and will not yield efficient algorithms with *meaningful* theoretical guarantees capturing the benefits of both contexts and side observations.

In this paper, we present the first study on adversarial linear contextual bandits with graph-structured side observations (or simply, graphical contextual bandits). Specifically, at each time step t, the adversary chooses the loss vector for each action in a finite set of K actions, and then a learning agent chooses from this K-action set after being presented with a d-dimensional context. After playing the chosen action, the agent not only incurs and observes the loss of the chosen action, but also observes losses of its neighboring action in the feedback graph G_t, where the losses are generated by the contexts and loss vectors under the linear payoff assumption (Agrawal and Goyal 2012). The goal of the agent is to minimize the regret, defined as the gap between the losses incurred by the agent and that of some suitable benchmark policy. Under mild conditions, we develop two algorithms for this problem with theoretical guarantees: i) `EXP3-LGC-U`, inspired by `EXP3-SET` (Alon et al. 2013, 2017) and `LinEXP3` (Neu and Olkhovskaya 2020); ii) `EXP3-LGC-IX`, inspired by `EXP3-IX` (Kocák et al. 2014) and `LinEXP3`. The contributions of the present work can be summarized as follows:

- We present and study a new bandit model, graphical contextual bandits, which jointly leverages two categories of the most common side information: contexts and side observations. This new model generalizes the original contextual bandits and graphical bandits, and turns out to be more delicate in describing real-world applications in social networks.
- Under mild assumptions, we propose the first algorithm, `EXP3-LGC-U`, for the general adversarial graphical contextual bandits. When the feedback graphs $\{G_t\}_{t=1}^T$ are undirected, we show that `EXP3-LGC-U` achieves a regret $\mathcal{O}(\sqrt{(K+\alpha(G)d)T\log K})$, where $\alpha(G)$ is the average *independence number* of $\{G_t\}_{t=1}^T$. In the directed graph setting, we show a slightly weaker result with a regret $\mathcal{O}(\sqrt{(K+\alpha(G)d)T}\log(KdT))$. When losses are non-negative, we develop the second algorithm, `EXP3-LGC-IX`, whose regret upper bound is $\mathcal{O}(\sqrt{\alpha(G)dT\log K\log(KT)})$ for both directed and undirected graph settings.
- In all regret upper bounds of our novel algorithms, the dependencies on d and T match exactly the best existing algorithm `RealLinEXP3` (Neu and Olkhovskaya 2020) for adversarial linear contextual bandits. Furthermore, the dependency on K of our proposed algorithms improves over `RealLinEXP3` as $\alpha(G) \leq K$ always holds, where the quantity $\alpha(G)$ matches the lower bound $\Omega(\sqrt{\alpha(G)T})$ for adversarial graphical bandits (Mannor and Shamir 2011). This comparison indicates that our proposed algorithms capture the benefits of both contexts and side observations.
- Numerical tests reveal the merits of the proposed model and algorithms over the state-of-the-art approaches.

The remainder of this paper is organized as follows. A brief literature review is presented in Section 2. Problem formulations and necessary assumptions for analysis are introduced in Section 3. The `EXP3-LGC-U` and `EXP3-LGC-IX` together with their analyses, are detailed in Sections 4 and 5, respectively. Finally, we conclude the paper in Section 7. Due to the page limitation, the proofs for the Claims, Theorems, and Corollaries are deferred to the appendix[1].

Notation. We use $\|x\|_2$ to denote the Euclidean norm of vector x; $\langle x, y\rangle$ stands for the inner product of x and y. We also define $\mathbb{E}_t[\cdot] = \mathbb{E}\left[\cdot \middle| \mathcal{F}_{t-1}\right]$ as the expectation given the filtration $\mathcal{F}_{t-1}$.

2 Related Work

Contextual bandits: Our paper studies a variant of adversarial contextual bandits, where adversarial contextual bandits were first investigated in Rakhlin and Sridharan (2016); Syrgkanis, Krishnamurthy, and Schapire (2016); Syrgkanis et al. (2016) for arbitrary class of policies without the linear payoff assumption. More relevant to our paper is Neu and Olkhovskaya (2020) that studied adversarial linear contextual bandits. Another category of contextual bandits is named as contextual bandits. For stochastic linear contextual bandits, Auer, Cesa-Bianchi, and Gentile (2002); Chu et al. (2011); Li et al. (2010); Abbasi-Yadkori, Pál, and Szepesvári (2011) provided `UCB`-based algorithms; Agrawal and Goyal (2013); Abeille and Lazaric (2017) designed and analyzed a generalization of Thompson sampling for the contextual setting. Stochastic contextual bandits in generalized linear models are studied in Valko et al. (2013); Filippi et al. (2010); Calandriello et al. (2019). Stochastic contextual bandits with arbitrary set of policies can be found in Langford and Zhang (2008); Dudik et al. (2011); Agarwal et al. (2014); Foster, Krishnamurthy, and Luo (2019); Foster et al. (2018); Foster and Rakhlin (2020). A neural net framework for stochastic contextual bandits with theoretical guarantees is proposed in Zhou, Li, and Gu (2020). Other interesting works include non-stationary contextual bandits (Luo et al. 2018; Chen et al. 2019), fairness in contextual bandits (Joseph et al. 2016; Chen et al. 2020), etc. We refer the readers to Zhou (2015) for a survey on contextual bandits.

Graphical bandits: If the contexts are not considered, our model will degenerate to Graphical bandits, which consider the side observations upon classical MAB. Graphical bansits were first proposed under the adversarial setting (Mannor and Shamir 2011). Performance for the model was then improved in a series of works (Alon et al. 2013; Kocák et al. 2014; Alon et al. 2015, 2017), with best performance matching the lower bound $\Omega(\sqrt{\alpha(G)T})$. Most existing algorithms for adversarial graphical bandits are based on the classical `EXP3`. Graphical bandits

[1] https://arxiv.org/abs/2012.05756

has also been considered in the stochastic setting: Caron et al. (2012) proposed a variant of UCB1; Buccapatnam, Eryilmaz, and Shroff (2014) improved the previous result via ϵ-greedy and UCB with a well-designed linear programming; Cohen, Hazan, and Koren (2016) developed an elimination-based algorithm that achieved the optimal regret; Thompson-sampling-based algorithms were recently proposed in Liu, Buccapatnam, and Shroff (2018); Liu, Zheng, and Shroff (2018). Other related works include graphical bandits with noisy observations (Kocák, Neu, and Valko 2016; Wu, György, and Szepesvári 2015), graphical bandits with switching costs (Arora, Marinov, and Mohri 2019; Rangi and Franceschetti 2019), graphical bandits with small-loss bound (Lee, Luo, and Zhang 2020; Lykouris, Sridharan, and Tardos 2018), etc. We refer the readers to Valko (2016) for a survey on graphical bandits.

Graphical contextual bandits: Recently, Singh et al. (2020) studied a stochastic variant of our model. UCB and linear programming (LP) based algorithms were proposed. The UCB based algorithm achieves a regret $\mathcal{O}(K \log T)$, whereas the LP based approach achieves a better regret $\mathcal{O}(\chi(G) \log T)$ with $\chi(G)$ denoting the dominant number.

3 Preliminaries

We consider an adversarial linear contextual bandit problem with graph-structured side observations between a *learning agent* with a finite action set $V := \{1, \dots, K\}$ and its *adversary*. At each time step $t = 1, 2, \dots, T$, the interaction steps between the agent and its adversary are repeated, which are described as follows. At the beginning of time step t, the feedback graph $G_t(V, \mathcal{E}_t)$ and a loss vector $\theta_{i,t} \in \mathbb{R}^d$ for each action $i \in V$ are chosen by the adversary arbitrarily, where G_t can be directed or undirected, V is the node set (the same as the action set V), and $\mathcal{E}_t$ is the edge set. Note that G_t and $\theta_{i,t}$ are *not* disclosed to the agent at this time. After observing a context $X_t \in \mathbb{R}^d$, the agent chooses an action $I_t \in V$ to play based on X_t, the previous interaction history, and possibly some randomness in the policy, and incurs the loss $\ell_t(X_t, I_t) = \langle X_t, \theta_{I_t,t} \rangle$. Unlike the recently proposed adversarial linear contextual bandits (Neu and Olkhovskaya 2020), where only the played action I_t discloses its loss $\ell_t(X_t, I_t)$, here we assume all losses in a subset $S_{I_t,t} \subseteq V$ are disclosed after I_t is played, where S_{I_t} contains I_t and its neighboring nodes in the feedback graph G_t. More formally, we have that $S_{i,t} := \{j \in V \big| i \xrightarrow{t} j \in \mathcal{E}_t \text{ or } j = i\}$, where $i \xrightarrow{t} j$ indicates an edge from node i to node j in a directed graph or an edge between i and j in an undirected graph at time t. These observations except for that of action I_t are called *side observations* in graphical bandits (Mannor and Shamir 2011). In addition, an *oracle* provides extra observations for all $i \in S_{I_t}$ (see Assumption 2 for details). Before proceeding to time step $t + 1$, the adversary discloses G_t to the agent.

Remark 1. The way the adversary discloses G_t in this paper is called the **uninformed** setting, where G_t is disclosed **after** the agent's decision making. Contrarily, a simpler setting from the agent's perspective is called the **informed** setting (Alon et al. 2013), where G_t is disclosed **before** the agent's decision making. The uninformed setting is the minimum requirement for our problem to capture the benefits of side observations (Cohen, Hazan, and Koren 2016).

Furthermore, we have the following assumptions for the above interaction steps.

Assumption 1 (i.i.d. contexts)**.** *The context $X_t \in \mathbb{R}^d$ is drawn from a distribution $\mathcal{D}$ independently from the choice of loss vectors and other contexts, where $\mathcal{D}$ is known by the agent in advance .*

Assumption 2 (extra observation oracle)**.** *Assume at each time step t, there exists an **oracle** that draws a context $\tilde{X}_t \in \mathbb{R}^d$ from $\mathcal{D}$ independently from the choice of loss vectors and other contexts, and discloses $\tilde{X}_t$ together with the losses $\tilde{l}_t(\tilde{X}_t, i) = \left\langle \tilde{X}_t, \theta_{i,t} \right\rangle$ for all $i \in S_{I_t,t}$ to the agent.*

Assumption 3 (nonoblivious adversary)**.** *The adversary can be **nonoblivious**, who is allowed to choose G_t and $\theta_{i,t}, \forall i \in V$ at time t according to arbitrary functions of the interaction history $\mathcal{F}_{t-1}$ before time step t. Here, $\mathcal{F}_t := \sigma(X_s, \tilde{X}_s, I_s, G_s, \{\ell_s(X_s, i)\}_{i \in S_s}, \{\tilde{\ell}_s(\tilde{X}_s, i)\}_{i \in S_s}, \forall s \leq t)$ is the filtration capturing the interaction history up to time step t.*

Remark 2. Assumption 1 is standard in the literature of adversarial contextual bandits (Neu and Olkhovskaya 2020; Rakhlin and Sridharan 2016; Syrgkanis, Krishnamurthy, and Schapire 2016; Syrgkanis et al. 2016). In fact, it has been shown that if both the contexts and loss vectors are chosen by the adversary, no algorithm can achieve a sublinear regret (Neu and Olkhovskaya 2020; Syrgkanis, Krishnamurthy, and Schapire 2016). The oracle in Assumption 2 is mainly adopted from the proof perspective, and its role will be clear in the analysis. In real-world applications, this oracle can be realized. Consider the viral marketing problem for an example. After the user and her/his followers complete the questionnaire and get the offers, they will probably purchase the products and leave online reviews after they experience those products. Then, the extra observations can be provided by those reviews. Assumption 3 indicates $\theta_{t,i}$ is a random vector with $\mathbb{E}_t[\theta_{i,t}] = \theta_{i,t}$, and a similar result holds for G_t. Note that a bandit problem with a nonoblivious adversary is harder than that with an oblivious adversary (Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2020) that chooses all loss vectors and feedback graphs before the start of the interactions.

The goal of the agent is to find a policy that minimizes its *expected cumulative loss*. Equivalently, we can adopt the *expected cumulative (pseudo) regret*, defined as the maximum gap between the expected cumulative loss incurred by the agent and that of a properly chosen policy set Π,

$$\mathcal{R}_T = \max_{\pi_T \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \left\langle X_t, \theta_{I_t,t} - \theta_{\pi_T(X_t),t} \right\rangle\right]$$

$$= \max_{\pi_T \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \in V} \left(\pi_t^a(i|X_t) - \pi_T(i|X_t)\right) \left\langle X_t, \theta_{i,t} \right\rangle\right],$$

where the expectation is taken over the randomness of the agent's policy and the contexts. It is widely acknowledged that competing with a policy that uniformly chooses the best action in each time step t while incurring an $o(T)$ regret is hopeless in the adversarial setting (Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2020). Thus, we adopt the fixed policy set Π proposed for adversarial linear contextual bandits (Neu and Olkhovskaya 2020),

$$\Pi := \{\pi_T \big| \text{all policies } \pi_T : \mathbb{R}^d \mapsto V\}, \tag{1}$$

where the decision given by $\pi_T \in \Pi$ only depends the current received context X_t. The best policy $\pi_T^* \in \Pi$ is the one that satisfies the following condition

$$\pi_T^*(i|x) = \mathbb{I}\{i = \arg\min_{j\in V} \sum_{t=1}^{T} \langle x, \mathbb{E}[\theta_{j,t}]\rangle\},\ \forall x \in \mathbb{R}^d,$$

which can be derived from the regret definition as shown in Neu and Olkhovskaya (2020).

Before presenting our algorithms, we will further introduce several common assumptions and definitions in linear contextual bandits and graphical bandits. We assume the context distribution $\mathcal{D}$ is supported on a bounded set with each $x \sim \mathcal{D}$ satisfying $\|x\|_2 \leq \sigma$ for some positive σ. Furthermore, we assume the covariance $\Sigma = \mathbb{E}[X_t X_t^\top]$ of $\mathcal{D}$ to be positive definite with its smallest eigenvalue being $\lambda_{\min} > 0$. As for the loss vector $\theta_{i,t}$, we assume that $\|\theta_{i,t}\|_2 \leq L$ for some positive L for all i, t. Additionally, the loss $\ell_t(x,t)$ is bounded in $[-1,1]$: $|\ell_t(x,i)| \leq 1$ for all $x \sim \mathcal{D}$, i, and t. We have the following graph-theoretic definition from Alon et al. (2013, 2017); Liu, Zheng, and Shroff (2018).

Definition 1 (Independence number)**.** *The cardinality of the maximum independent set of a graph G_t is defined as the* ***independence number*** *and denoted by $\alpha(G_t)$, where an independence set of $G_t = (V_t, \mathcal{E}_t)$ is any subset $V_t' \in V_t$ such that no two nodes $i, j \in V_t'$ are connected by an edge in $\mathcal{E}_t$. Note that $\alpha(G_t) \leq K$ in general. Without ambiguity, we use $\alpha(G) := \frac{1}{T}\sum_{t=1}^{T}\alpha(G_t)$ to denote the average* independence number *of the feedback graphs $\{G_t\}_{t=1}^T$ in remainder of this paper.*

4 The EXP3-LGC-U Algorithm

In this section, we introduce our first simple yet efficient algorithm, EXP3-LGC-U, for both directed and undirected feedback graphs, which is the abbreviation for "**EXP3** for **L**inear **G**raphical **C**ontextual bandits with **U**niform exploration". Detailed steps of EXP3-LGC-U are presented in Algorithm 1. The upper bounds for the regret of EXP3-LGC-U are developed in Section 4.1. We further discuss our theoretical findings on EXP3-LGC-U in Section 4.2.

The core of our algorithm, similar to many other algorithms for adversarial bandits, is designing an appropriate estimator of each loss vector and using those estimators to define a proper policy. Following the EXP3-based algorithms, we apply an exponentially weighted method and play an action i with probability proportional to $\exp(-\eta\sum_{s=1}^{t-1}\langle X_t, \hat{\theta}_{i,s}\rangle)$ (see Eq. (2)) at time step t, where η is the learning rate. More precisely, a uniform exploration γ is needed for the probability distribution of drawing action (see Eq. (3)). The uniform exploration is to control the variance of the loss vector estimators, which is a key step in our analysis. At this point, the key remaining question is how to design a reasonable estimator for each loss vector $\theta_{i,t}$. The answer can be found in Eq. (4), which takes advantage of both the original observations and the extra observations from the oracle. Similar to EXP3-SET, our algorithm uses importance sampling to construct the loss vector estimator $\hat{\theta}_{i,t}$ with controlled variance. The term $q_t(i|X_t)$ in the denominator in Eq. (4) indicates the probability of observing the loss of action i at time t, which is simply the sum of all $\pi_t^a(j|X_t)$ for all j that is connected to i at time t. The reason we use $\tilde{\ell}(\tilde{X}_t, i)$ and $\tilde{X}_t$ instead of $\ell(\tilde{X}_t, i)$ and X_t in constructing loss vector estimator $\hat{\theta}_{i,t}$ can be partly interpreted in the following two claims.

Algorithm 1 EXP3-LGC-U

```
Input: Learning rate $\eta > 0$, uniform exploration rate $\gamma \in (0,1)$, covariance $\Sigma$, and action set $V$.
For $t = 1, \ldots, T$, do:
  1. Feedback graph $G_t$ and loss vectors $\{\theta_{i,t}\}_{i\in V}$ are generated but not disclosed.
  2. Observe $X_t \sim \mathcal{D}$, and for all $i \in V$, set
       $w_t(X_t, i) = \exp\left(-\eta\sum_{s=1}^{t-1}\left\langle X_t, \hat{\theta}_{i,s}\right\rangle\right).$   (2)
  3. Play action $I_t$ drawn according to distribution $\pi_t^a(X_t) := (\pi_t^a(1|X_t), \ldots, \pi_t^a(K|X_t))$, where
       $\pi_t^a(i|X_t) = (1-\gamma)\frac{w_t(X_t,i)}{\sum_{j\in V} w_t(X_t,j)} + \frac{\gamma}{K}.$   (3)
  4. Observe pairs $(i, \ell_t(X_t, i))$ for all $i \in S_{I_t,t}$, and disclose feedback graph $G_t$.
  5. Extra observation oracle: observe $\tilde{X}_t \sim \mathcal{D}$ and pairs $(i, \tilde{\ell}_t(\tilde{X}_t, i))$ for all $i \in S_{I_t,t}$.
  6. For each $i \in V$, estimate the loss vector $\theta_{i,t}$ as
       $\hat{\theta}_{i,t} = \frac{\mathbb{I}\{i \in S_{I_t,t}\}}{q_t(i|X_t)}\Sigma^{-1}\tilde{X}_t\tilde{\ell}_t(\tilde{X}_t, i),$   (4)
     where $q_t(i|X_t) = \pi_t^a(i|X_t) + \sum_{j:j\xrightarrow{t}i}\pi_t^a(j|X_t)$.
End For
```

Claim 1. *The estimator $\hat{\theta}_{i,t}$ of the loss vector $\theta_{i,t}$ in Eq.* (4) *is an unbiased estimator given the interaction history $\mathcal{F}_{t-1}$ and X_t, for each $i \in V$ and t, i.e., $\mathbb{E}_t\left[\hat{\theta}_{i,t}\middle| X_t\right] = \theta_{i,t}$.*

It is straightforward to show that the estimator $\hat{\theta}_{i,t}$ in Eq. (4) is unbiased w.r.t. $\mathbb{E}_t[\cdot]$ and $\mathbb{E}[\cdot]$ by applying the law of total expectation. However, if we use X_t and $\ell(X_t, i)$ to construct $\hat{\theta}_{i,t}$ in Eq. (4), it will only be unbiased w.r.t. $\mathbb{E}_t[\cdot]$ and $\mathbb{E}[\cdot]$, but not $\mathbb{E}_t[\cdot|X_t]$. This observation turns out to be essential in our analysis, which leads to the following immediate result of Claim 1.

Claim 2. *Let $\pi_T : \mathbb{R}^d \mapsto V$ be any policy in Π and $\hat{\theta}_{i,t}$ follows Eq. (4). Suppose π_t^a is determined by $\mathcal{F}_{t-1}$ and X_t, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V}\left(\pi_t^a(i|X_t) - \pi_T(i|X_t)\right)\langle X_t, \theta_{i,t}\rangle\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V}\left(\pi_t^a(i|X_t) - \pi_T(i|X_t)\right)\left\langle X_t, \hat{\theta}_{i,t}\right\rangle\right]. \quad (5)$$

Remark 3. The advantages and properties of Claim 2 are summarized as following. i) By applying the policy produced by EXP3-LGC-U and the best policy in the fixed policy set Π in Eq. (1), the term in the right hand side of Eq. (5) is exactly the regret $\mathcal{R}_T$ of EXP3-LGC-U. Given this property, the known loss vector estimate $\hat{\theta}_{i,t}$, instead of the unknown true loss vector $\theta_{i,t}$, can be applied directly to our analysis of the regret. ii) Claim 2 is not confined to EXP3-LGC-U and can be applied to other loss vector estimators that adopt different construction methods and any other benchmark policy, as long as Claim 1 is satisfied. iii) Based on Claim 2, some techniques in proving classical EXP3 can be utilized in our analysis of the regret.

Remark 4. Claim 2 exhibits several differences between adversarial contextual bandits and classical adversarial MAB. First, the benchmark policy $\pi_T(\cdot|X_t)$ depends on the contexts in adversarial contextual bandits, while the benchmark policy is the best fixed action in hindsight in classical adversarial MAB. Second, consider the regret definition of classical adversarial MAB, $\mathcal{R}_T^{\text{MAB}} = \max_{j\in V}\mathbb{E}\left[\sum_{t=1}^{T}(\sum_{i\in V}\pi_t^{a,\text{MAB}}(i)\ell_{i,t}) - \ell_{j,t}\right]$, where $\pi_t^{a,\text{MAB}}(i)$ is the policy produced by an EXP3-based algorithm and $\ell_{i,t}$ is the loss for action i at time step t. Since no context exists here, it is natural to design an estimator $\hat{\ell}_{i,t}$ of $\ell_{i,t}$ that is unbiased w.r.t. $\mathbb{E}_t[\cdot]$, and a similar result as Claim 2 can be proved. However, with the contexts, if the loss vector estimators are only unbiased w.r.t. $\mathbb{E}_t[\cdot]$ rather than $\mathbb{E}_t[\cdot|\, X_t]$, Claim 2 will not hold as shown in the proof of Claim 2 in Appendix A.2.

Remarks 3 and 4 explain the need of adopting the extra observation oracle in EXP3-LGC-U and the way the loss vector estimator $\hat{\theta}_{i,t}$ is constructed.

4.1 Regret Analysis for EXP3-LGC-U

Our main theoretical justification for the performance of EXP3-LGC-U summarized in Theorem 1.

Theorem 1. *For any positive $\eta \in (0,1)$, choosing $\gamma = \eta K\sigma^2/\lambda_{min}$, the expected cumulative regret of EXP3-LGC-U satisfies:*

$$\mathcal{R}_t \le \frac{\log K}{\eta} + \frac{2\eta K\sigma^2}{\lambda_{min}}T + \eta d\sum_{t=1}^{T}\mathbb{E}\left[Q_t\right],$$

where $Q_t = \alpha(G_t)$ if G_t is undirected, and $Q_t = 4\alpha(G_t)\log(4K^2/(\alpha(G_t)\gamma))$ if G_t is directed.

The proof of Theorem 1 is mainly based on the following Lemma 1, which is established on Claim 2.

Lemma 1. *Supposing $\left|\eta\left\langle X_t, \hat{\theta}_{i,t}\right\rangle\right| \le 1$, the expected cumulative regret of EXP3-LGC-U satisfies*

$$\mathcal{R}_T \le \frac{\log K}{\eta} + 2\gamma T + \eta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V}\pi_t^a(i|X_t)\left\langle X_t, \hat{\theta}_{i,t}\right\rangle^2\right]. \quad (6)$$

The proof of Lemma 1 is detailed in Appendix A.3. The last term in the right side of Eq. (6) can be further bounded using graph-theoretic results in Alon et al. (2017, Lemma 10) and Alon et al. (2015, Lemma 5), which are restated in Appendix A.4.

Remark 5. According to Eq. (13) in the proof of Theorem 1 in Appendix A.4, if the extra observation oracle is not adopted, we will have a higher-order term $\mathbb{E}\left[X_t^\top\Sigma^{-1}X_tX_t^\top\Sigma^{-1}X_t\right]$. In general, it is hard to specify the relationship between this term and the dimension of contexts d. This explains why we adopt the oracle in the algorithm.

We have the following two corollaries based on Theorem 1, where the notations follow Alon et al. (2013, 2017).

Corollary 1. *For the undirected graph setting, if $\alpha(G_t) \le \alpha_t$ for $t = 1,\ldots,T$, then setting $\eta = \sqrt{\log K/\left(2K\sigma^2T/\lambda_{min} + d\sum_{t=1}^{T}\alpha_t\right)}$ gives $\mathcal{R}_T = \mathcal{O}\left(\sqrt{\left(2K\sigma^2T/\lambda_{min} + d\sum_{t=1}^{T}\alpha_t\right)\log K}\right)$.*

Corollary 2. *For the directed graph setting, if $\alpha(G_t) \le \alpha_t$ for $t = 1,\ldots,T$, and supposing that T is large enough so that $\log(1/\gamma) \ge 1$, then setting $\eta = \left(2K\sigma^2T/\lambda_{min} + 4d\sum_{t=1}^{T}\alpha_t\right)^{-\frac{1}{2}}$ gives $\mathcal{R}_T = \mathcal{O}\left(\sqrt{2K\sigma^2T/\lambda_{min} + 4d\sum_{t=1}^{T}\alpha_t}\log(KdT)\right)$.*

4.2 Discussion

Corollaries 1 and 2 reveal that by properly choosing the learning rate η and the uniform exploration rate γ, the regret of EXP3-LGC-U can be upper bounded by $\mathcal{O}(\sqrt{(K+\alpha(G)d)T\log K})$ in the undirected graph setting, and $\mathcal{O}(\sqrt{(K+\alpha(G)d)T}\log(KdT))$ in the directed graph setting. Compared with state-of-the-art algorithms for adversarial linear contextual bandits, EXP3-LGC-U has tighter regret upper bounds in the extreme case when the feedback graph G_t is a fixed edgeless graph ($\alpha(G) = K$), as Neu and Olkhovskaya (2020) shows $\mathcal{O}(5T^{2/3}(Kd\log K)^{1/3})$ for RobustLinEXP3 and $\mathcal{O}(4\sqrt{T} + \sqrt{dKT\log K}(3 + \sqrt{\log T}))$ for RealLinEXP3. It is easily verified that the dependencies on d and T in the regrets of EXP3-LGC-U match with the best existing algorithm RealLinEXP3. Furthermore, the dependence on K of EXP3-LGC-U is matching with the lower bound $\Omega(\sqrt{\alpha(G)T})$ for graphical bandits (Mannor and Shamir 2011), which improves

```
Algorithm 2 EXP3-LGC-IX
Parameters: Learning rate $\eta_t > 0$, implicit exploration rate $\beta_t \in (0,1)$, and covariance $\Sigma$,and action set $V$.
For $t = 1, \ldots, T$, do:
  1. Feedback graph $G_t$ and loss vectors $\{\theta_{i,t}\}_{i \in V}$ are generated but not disclosed.
  2. Observe $X_t \sim \mathcal{D}$, and play action $I_t$ drawn according to distribution $\pi_t^a(X_t) := (\pi_t^a(1|X_t), \ldots, \pi_t^a(K|X_t))$ with
       $$\pi_t^a(i|X_t) = \frac{w_t(X_t,i)}{\sum_{j \in V} w_t(X_t,j)}, \quad (7)$$
     where $w_t(X_t,i) = \frac{1}{K} \exp\left(-\eta_t \sum_{s=1}^{t-1} \left\langle X_t, \hat{\theta}_{i,s} \right\rangle\right)$.
  3. Observe pairs $(i, \ell_t(X_t,i))$ for all $i \in S_{I_t,t}$, disclose feedback graph $G_t$.
  4. Extra observation oracle: observe $\tilde{X}_t \sim \mathcal{D}$ and pairs $(i, \tilde{\ell}_t(\tilde{X}_t,i))$ for all $i \in S_{I_t,t}$.
  5. For each $i \in V$, estimate the loss vector $\theta_{i,t}$ as
       $$\hat{\theta}_{i,t} = \frac{\mathbb{I}\{i \in S_{I_t,t}\}}{q_t(i|X_t) + \beta_t} \Sigma^{-1} \tilde{X}_t \tilde{\ell}_t(\tilde{X}_t,i), \quad (8)$$
     where $q_t(i|X_t) = \pi_t^a(i|X_t) + \sum_{j:j \xrightarrow{t} i} \pi_t^a(j|X_t)$.
End For
```

over that of RealLinEXP3 in general cases. Moreover, our result is also better than algorithms designed for adversarial contextual bandits with arbitrary class of policies (Rakhlin and Sridharan 2016; Syrgkanis, Krishnamurthy, and Schapire 2016; Syrgkanis et al. 2016), which are not capable of guaranteeing an $\mathcal{O}(\sqrt{T})$ regret.

In addition, Neu and Olkhovskaya (2020) is different from ours in the following respects: i) loss vector estimator construction, and ii) proof techniques. First, the estimator in Neu and Olkhovskaya (2020) is only unbiased w.r.t. $\mathbb{E}_t[\cdot]$ rather than $\mathbb{E}_t[\cdot| X_t]$. Second, their proof is conducted on an auxiliary online learning problem for a fixed context X_0 with K actions (See Neu and Olkhovskaya (2020, Lemmas 3 and 4) for details).

5 The EXP3-LGC-IX Algorithm

In this section, we present another efficient algorithm, EXP3-LGC-IX, for a special class of problems when the support of $\theta_{i,t}$ and X_t is non-negative, and elements of X_t are independent. The motivation for such a setting still comes from the viral marketing problem. Suppose the agent has a questionnaire (context) of some product, which contains true/false questions that are positively weighted. In this case, the answers of users (loss vectors) will be vectors that contain only $0/1$ entries. Under the linear payoff assumption, the loss is non-negative. EXP3-LGC-IX, which is the abbreviation for "**EXP3** for **L**inear **G**raphical **C**ontextual bandits with **I**mplicit e**X**ploration", has the same regret upper bound for both directed and undirected graph settings, as shown in Section 5.1.

Algorithm 2 shows the detailed steps of EXP3-LGC-IX, which follows the method of classical EXP3 and is similar to EXP3-LGC-U. The main differences between EXP3-LGC-IX and EXP3-LGC-U are as follows. First, no explicit uniform exploration mixes with the probability distribution of drawing action (see Eq. (7)). In this case, for EXP3-LGC-U without uniform exploration, only a worse regret upper bound that contains $mas(G)$ rather than $\alpha(G)$ can be proved in the directed graph setting, where $mas(G)$ is the average *maximum acyclic subgraphs number* and $mas(G) \geq \alpha(G)$. This result could be obtained by simply removing the uniform exploration part in the proof of EXP3-LGC-U and substituting Lemma 3 with Alon et al. (2017, Lemma 10). Second, biased loss vector estimator is adopted (see Eq. (8)). Similar to EXP3-IX, this biased estimator ensures that the loss estimator satisfies the following claim, which turns out to be essential for our analysis.

Claim 3. *The estimator $\hat{\theta}_{i,t}$ of the loss vector $\theta_{i,t}$ for each $i \in V$ and t satisfies*

$$\mathbb{E}_t\left[\sum_{i \in V} \pi_t^a(i|X_t) \left\langle X_t, \hat{\theta}_{i,t} \right\rangle \middle| X_t\right] = \sum_{i \in V} \pi_t^a(i|X_t) \left\langle X_t, \theta_{i,t} \right\rangle$$
$$- \beta_t \sum_{i \in V} \frac{\pi_t^a(i|X_t)}{q_t(i|X_t) + \beta_t} \left\langle X_t, \theta_{i,t} \right\rangle. \quad (9)$$

Remark 6. Claim 3 indicates the loss estimators in EXP3-LGC-IX are optimistic. The bias incurred by EXP3-LGC-IX can be directly controlled by the implicit exploration rate β_t. This kind of implicit exploration actually has similar effect in controlling the variance of the loss estimators as explicit exploration (e.g., uniform exploration), though the approach is different. Notice that Claim 3 does not hold if there is no extra observation oracle (see the proof in Appendix B.1 for details), which further demonstrates the necessity of the oracle.

5.1 Regret Analysis for EXP3-LGC-IX

The upper bound on the regret of EXP3-LGC-IX follows Theorem 2, where the proof of Theorem 2 is deferred to Appendix B.2. Notice that a similar higher-order term as that in Remark 5 appear in the proof of Theorem 2, if the extra observation oracle is not adopted.

Theorem 2. *Setting $\beta_t = \sqrt{\log K/(K + \sum_{s=1}^{t-1} Q_s)}$ and $\eta_t = \sqrt{\log K/(dK + d\sum_{s=1}^{t-1} Q_s)}$, the expected regret of* EXP3-LGC-IX *satisfies:*

$$\mathcal{R}_T \leq 2(1+\sqrt{d})\mathbb{E}\left[\sqrt{\left(K + \sum_{t=1}^{T} Q_t\right) \log K}\right] \quad (10)$$

for both directed and undirected graph settings, where $Q_t = 2\alpha(G_t) \log\left(1 + \frac{\lceil K^2/\beta_t \rceil + K}{\alpha(G_t)}\right) + 2$.

Based on Theorem 2, we have the following corollary.

Corollary 3. *Suppose $\alpha(G_t) \leq \alpha_t$ for $t = 1, \ldots T$, the regret of* EXP3-LGC-IX *satisfies $\mathcal{R}_T = \mathcal{O}(\sqrt{\sum_{t=1}^{T} \alpha_t d \log K \log(KT)})$, for both directed and undirected graph settings.*

Corollary 3 reveals that by adopting the learning rate η_t and the implicit exploration rate β_t adaptively, the regret of EXP3-LGC-IX can be upper bounded by $\mathcal{O}(\sqrt{\alpha(G)dT\log K\log(KT)})$ for both directed and undirected graph settings. This result indicates that EXP3-LGC-IX captures the benefits of both contexts and side observations, as discussed in Section 4.2. The EXP3-LGC-IX algorithm cannot handle negative losses due to the following two reasons. First, if the losses are negative, Claim 3 does not hold. Second, although we can flip the sign of β_t according to the sign of the loss vector to guarantee the optimism of the loss estimator, the graph-theoretic result (e.g., Kocák et al. (2014, Lemma 2)) cannot be applied as β_t is required to be positive.

6 Numerical Results

We conduct the numerical tests on synthetic data to demonstrate the efficiency of the novel EXP3-LGC-U and EXP3-LGC-IX algorithms.

We consider a setting of $K = 10$ actions, with $d = 10$ dimensional contexts observed iteratively on a $T = 10^5$ time horizon. Each coordinate of context X_t (or $\tilde{X}_t$) is generated i.i.d. from the Bernoulli distribution with support $\{0, 1/\sqrt{d}\}$ and $p = 0.5$, where the covariance of X_t is $I_d/(4d)$, and I_d is the identity matrix of size $d \times d$. The loss vectors are generated with a sudden change. Specially, for $t \in [1, 50000]$, each coordinate of $\theta_{i,t}$ are set to be $\theta_{i,t}(j) = 0.1i|\cos t|/\sqrt{d}$, whereas $\theta_{i,t}(j) = 0.05i|\sin t|/\sqrt{d}$ for the remaining time steps, for all $j = 1, \dots, d$. We consider the time-invariant and undirected feedback graph structure for the purpose of performance validation. As depicted in Figure 1, the feedback graph consists of a 9-vertex complete graph and one isolated vertex, where the independence number $\alpha(G) = 2$.

We compare EXP3-LGC-U and EXP3-LGC-IX with RobustLinEXP3 from Neu and Olkhovskaya (2020). We also let EXP3-LGC-U* and EXP3-LGC-IX* denote the proposed algorithms without relying on side observations. The parameters of EXP3-LGC-U and EXP3-LGC-IX are chosen according to Corollary 1 and Theorem 2, respectively. For EXP3-LGC-U* and EXP3-LGC-IX*, the parameter selection methods are identical as before, except for setting $\alpha(G) = K$. The parameters of RobustLinEXP3 are tuned exactly the same as those in Neu and Olkhovskaya (2020).

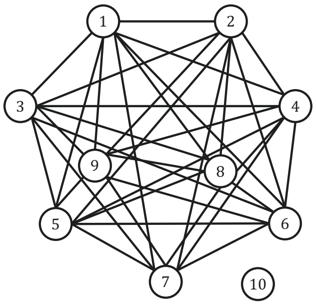

Figure 1: Feedback graph structure for the numerical tests.

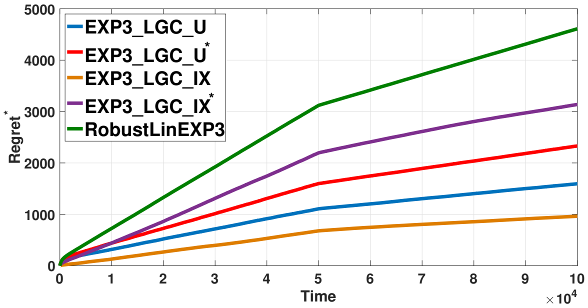

Figure 2: Regret[2] comparison of different algorithms on synthetic dataset over 100 independent trials.

Figure 2 presents the expected cumulative regret[2], where the results are averaged over 100 independent trials. We find that EXP3-LGC-U and EXP3-LGC-IX significantly outperform the baseline algorithms (RobustLinEXP3, EXP3-LGC-U*, and EXP3-LGC-IX*), which is consistent with theoretical guarantees in Theorems 1 and 2. Besides, even if there is no side observation, our proposed algorithms are also better than RobustLinEXP3 (see comparison among EXP3-LGC-U*, EXP3-LGC-IX*, and RobustLinEXP3 in Figure 2).

7 Conclusion

We introduce a new MAB formulation – adversarial graphical contextual bandits – which leverage both contexts and side observations. Two efficient algorithms, EXP3-LGC-U and EXP3-LGC-IX, are proposed, with EXP3-LGC-IX for a special class of problems and EXP3-LGC-U for more general cases. Under mild assumptions, it is analytically demonstrated that the proposed algorithms achieve the regret $\widetilde{\mathcal{O}}(\sqrt{\alpha(G)dT})$ for both directed and undirected graph settings.

Several interesting questions are left open for future work. One challenging problem lies in providing a tight lower bound for adversarial linear graphical contextual bandits. Another promising direction for follow-up work is studying the small-loss bound for graphical contextual bandits.

Acknowledgements

Lingda Wang and Zhizhen Zhao are supported in part by Alfred P. Sloan Foundation. Bingcong Li and Georgios B. Giannakis gratefully acknowledge the support from NSF grants 1711471, and 1901134. Huozhi Zhou and Lav R. Varshney are funded in part by the IBM-Illinois Center for Cognitive Computing Systems Research (C3SR), a research collaboration as part of the IBM AI Horizons Network.

[2] The regret before time step T is actually the difference of accumulative losses between the algorithm and the benchmark policy for the horizon T, not the true regret defined for a horizon less than T.

References

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

Abeille, M.; and Lazaric, A. 2017. Linear Thompson sampling revisited. *Electronic Journal of Statistics* 11(2): 5165–5197.

Agarwal, A.; Hsu, D.; Kale, S.; Langford, J.; Li, L.; and Schapire, R. 2014. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 1638–1646.

Agrawal, S.; and Goyal, N. 2012. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 1–26.

Agrawal, S.; and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135.

Alon, N.; Cesa-Bianchi, N.; Dekel, O.; and Koren, T. 2015. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, 1–13.

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*, 1610–1618.

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

Audibert, J.-Y.; and Bubeck, S. 2009. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, 217–226.

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

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1): 48–77.

Auer, P.; Cesa-Bianchi, N.; and Gentile, C. 2002. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences* 64(1): 48–75.

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.

Buccapatnam, S.; Eryilmaz, A.; and Shroff, N. B. 2014. Stochastic bandits with side observations on networks. In *ACM International Conference on Measurement and Modeling of Computer Systems*, 289–300.

Calandriello, D.; Carratino, L.; Lazaric, A.; Valko, M.; and Rosasco, L. 2019. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, 533–557.

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

Chen, Y.; Cuellar, A.; Luo, H.; Modi, J.; Nemlekar, H.; and Nikolaidis, S. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 181–190.

Chen, Y.; Lee, C.-W.; Luo, H.; and Wei, C.-Y. 2019. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, 696–726.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 208–214.

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

Dudik, M.; Hsu, D.; Kale, S.; Karampatziakis, N.; Langford, J.; Reyzin, L.; and Zhang, T. 2011. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, 169–178.

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

Foster, D. J.; Agarwal, A.; Dudik, M.; Luo, H.; and Schapire, R. 2018. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, 1539–1548.

Foster, D. J.; Krishnamurthy, A.; and Luo, H. 2019. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, 14714–14725.

Foster, D. J.; and Rakhlin, A. 2020. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926* .

Jiang, C. 2015. *Online advertisements and multi-armed bandits*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 325–333.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, 199–213. Springer.

Kocák, T.; Neu, G.; and Valko, M. 2016. Online learning with noisy side observations. In *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*, 613–621.

Kveton, B.; Szepesvari, C.; Wen, Z.; and Ashkan, A. 2015. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, 767–776.

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

Langford, J.; and Zhang, T. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, 817–824.

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

Lee, C.-W.; Luo, H.; and Zhang, M. 2020. A closer look at small-loss bounds for bandits with graph feedback. In *Conference on Learning Theory*, 1–49.

Li, B.; Chen, T.; and Giannakis, G. B. 2019. Bandit online learning with unknown delays. In *International Conference on Artificial Intelligence and Statistics*, 993–1002.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, 661–670.

Liu, F.; Buccapatnam, S.; and Shroff, N. 2018. Information directed sampling for stochastic bandits with graph feedback. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Lobel, I.; Sadler, E.; and Varshney, L. R. 2017. Customer referral incentives and social media. *Management Science* 63(10): 3514–3529.

Luo, H.; Wei, C.-Y.; Agarwal, A.; and Langford, J. 2018. Efficient contextual bandits in non-stationary worlds. In *Conference on Learning Theory*, 1739–1776.

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

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*, 684–692.

Neu, G.; and Olkhovskaya, J. 2020. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, 1–20.

Rakhlin, A.; and Sridharan, K. 2016. BISTRO: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning*, 1977–1985.

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

Singh, R.; Liu, F.; Liu, X.; and Shroff, N. 2020. Contextual bandits with side-observations. *arXiv preprint arXiv:2006.03951* .

Syrgkanis, V.; Krishnamurthy, A.; and Schapire, R. 2016. Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, 2159–2168.

Syrgkanis, V.; Luo, H.; Krishnamurthy, A.; and Schapire, R. E. 2016. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems*, 3135–3143.

Tewari, A.; and Murphy, S. A. 2017. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 495–517. Springer.

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.

Valko, M. 2016. *Bandits on graphs and structures*. Ph.D. thesis, École Normale Supérieure de Cachan.

Valko, M.; Korda, N.; Munos, R.; Flaounas, I.; and Cristianini, N. 2013. Finite-time analysis of kernelised contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, 654–663.

Wang, L.; Zhou, H.; Li, B.; Varshney, L. R.; and Zhao, Z. 2019. Nearly optimal algorithms for piecewise-stationary cascading bandits. *arXiv preprint arXiv:1909.05886* .

Wu, Y.; György, A.; and Szepesvári, C. 2015. Online learning with Gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, 1360–1368.

Zhou, D.; Li, L.; and Gu, Q. 2020. Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, 11492–11502.

Zhou, H.; Wang, L.; Varshney, L. R.; and Lim, E.-P. 2020. A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Zhou, L. 2015. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326* .