Stochastic Bandits with Graph Feedback in Non-Stationary Environments

Shiyin Lu,[1] Yao Hu,[2] Lijun Zhang[1,*]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2] YouKu Cognitive and Intelligent Lab, Alibaba Group, Beijing 100102, China
{lusy, zhanglj}@lamda.nju.edu.cn, yaoohu@alibaba-inc.com

Abstract

We study a variant of stochastic bandits where the feedback model is specified by a graph. In this setting, after playing an arm, one can observe rewards of not only the played arm but also other arms that are adjacent to the played arm in the graph. Most of the existing work assumes the reward distributions are stationary over time, which, however, is often violated in common scenarios such as recommendation systems and online advertising. To address this limitation, we study stochastic bandits with graph feedback in non-stationary environments and propose algorithms with graph-dependent dynamic regret bounds. When the number of reward distribution changes L is known in advance, one of our algorithms achieves an $\tilde{O}(\sqrt{\alpha LT})$ dynamic regret bound. We also develop an adaptive algorithm that can adapt to unknown L and attain an $\tilde{O}(\sqrt{\theta LT})$ dynamic regret. Here, α and θ are some graph-dependent quantities and T is the time horizon.

Introduction

Stochastic bandits are a powerful learning paradigm for sequential decision-making under uncertainty and have been applied in a variety of real-world scenarios such as online advertising (Chen, Wang, and Yuan 2013), news recommendation (Li et al. 2010), and social networks (Bnaya et al. 2013). A canonical model for studying this paradigm is the stochastic multi-armed bandits (MAB). In each round of MAB, a learner has to choose one of K arms to play. After playing an arm, the learner observes a stochastic reward drawn from the distribution associated with the played arm, while rewards of other arms remain unknown. The learner's goal is to minimize the regret, which is the difference between the cumulative reward of arms chosen by the learner and that of the best arm in hindsight. To accomplish this goal, the learner needs to balance the trade-off between exploration (choosing less played arms to gain more information) and exploitation (selecting seemingly optimal arms to accumulate more reward). Since the pioneering work of Thompson (1933) and Robbins (1952), the stochastic MAB has been widely studied and it is known that the minimax regret bound is $\Theta(K \log T)$ (Lai and Robbins 1985; Auer, Cesa-Bianchi, and Fischer 2002).

While this bound is logarithmic in T, it becomes vacuous when K is large, revealing that MAB is not suitable for applications with too many arms. Another limitation of MAB is that the learner is assumed to observe reward of only the chosen arm, which is too pessimistic as in some applications including recommendation systems and online advertising, side observations on rewards of other arms are available (Alon et al. 2017). To address these limitations, Mannor and Shamir (2011) and Caron et al. (2012) introduced a variant of MAB—bandits with graph feedback.[1] In this setting, there exists an undirected graph $G = (V, E)$, where V is the vertex set consisting of all arms, and E is the edge set. An edge $e = (u, v)$ in E indicates that after playing arm u, the learner can observe reward of not only arm u but also arm v, and vice versa. For stochastic bandits with graph feedback, Caron et al. (2012) proposed algorithms that enjoy regret bounds of $O(\theta \log T)$, where $\theta \leq K$ is the clique covering number of the feedback graph G, and can be much smaller than K for benign graphs.

Stochastic bandits with graph feedback were further extensively studied by a line of research (Buccapatnam, Eryilmaz, and Shroff 2014; Cohen, Hazan, and Koren 2016; Tossou, Dimitrakakis, and Dubhashi 2017; Liu, Buccapatnam, and Shroff 2018; Liu, Zheng, and Shroff 2018; Hu, Mehta, and Pan 2019; Lykouris, Tardos, and Wali 2019). However, most of the existing work assumes the reward distribution of each arm is stationary over time and thus does not apply to non-stationary rewards arising in the aforementioned real-world scenarios. For example, in recommendation systems, users' preference changes with time (Min and Han 2005). In online advertising, the click-through-rate of an advertisement is also time-variant (Zeng et al. 2016). Till now, we have very limited knowledge on stochastic bandits with graph feedback in non-stationary environments. One result was given by Alami (2019), who studied a rather limited setting of this problem, where the reward of each arm follows the Bernoulli distribution and in each round with a fixed probability, reward distributions of all arms change simultaneously. While Alami (2019) proposed a Thompson sampling algorithm for this setting, there is no theoretical guarantee of the proposed algorithm. In this paper, we inves-

*Lijun Zhang is the corresponding author.

[1] Mannor and Shamir (2011) considered adversarial bandits, while Caron et al. (2012) studied stochastic bandits.

tigate this problem under a much more general setting in the sense that the assumption on reward distributions is relaxed to allow any distribution with bounded support, and the reward distribution of each arm can change in an arbitrary manner. We adopt the dynamic regret (Jadbabaie et al. 2015; Auer, Gajane, and Ortner 2019), which compares the learner against an omniscient policy that in each round chooses the arm with the maximal mean reward, as performance metric, and develop three algorithms with different flavors.

As a warm-up, our first algorithm is a variant of the UCB-NE algorithm (Hu, Mehta, and Pan 2019) using the sliding-window mean estimator (Garivier and Moulines 2011), for which we prove an $\tilde{O}(\theta\sqrt{LT})$ dynamic regret bound.[2] While this algorithm is simple to understand, its regret bound is sub-optimal with respect to θ in the worst case ($\theta = K$), in light of the state-of-the-art $\tilde{O}(\sqrt{KLT})$ dynamic regret bound (Auer et al. 2002; Allesiardo, Féraud, and Maillard 2017; Auer, Gajane, and Ortner 2019) for the multi-armed bandits setting. To overcome this limitation, we then develop the second algorithm called SEASIDE. Built upon the successive elimination framework (Even-Dar, Mannor, and Mansour 2006), SEASIDE exploits the structure of the feedback graph to reduce the exploration cost, and randomly restarts itself to handle non-stationary environments. Theoretical analysis shows that SEASIDE attains an $\tilde{O}(\sqrt{\alpha LT})$ dynamic regret, where $\alpha \leq K$ is the independence number of the feedback graph and is not more than the clique covering number θ. A common issue of our first and second algorithms is that they need to know the number of reward distribution changes L in advance. Without such prior knowledge, their regret bounds scale with L instead of $\sqrt{L}$. In the setting of multi-armed bandits, this issue was solved by a recent milestone, the AdSwitch algorithm (Auer, Gajane, and Ortner 2019), the basic idea of which is to actively detect changes of reward distributions and restart itself once a change is detected. We extend AdSwitch to the graph feedback setting by designing a novel sampling scheme and a corresponding arm selection strategy that can utilize graph feedback. The resulting algorithm, called AdSwitch for Graph feedback (ASG), does not require prior knowledge of L and enjoys an $\tilde{O}(\sqrt{\theta LT})$ dynamic regret bound.

The regret bounds of both SEASIDE and ASG are optimal in terms of α, θ, L, and T up to logarithmic factors, since for any value of θ we can always construct a graph with $\theta = \alpha$ and the matching lower bounds of $\Omega(\sqrt{\alpha T})$ and $\Omega(\sqrt{LT})$ have been established (Mannor and Shamir 2011; Garivier and Moulines 2011; Zhou et al. 2020).

Related Work

In the pioneering papers (Kocsis, Szepesvári, and Willemson 2006; Hartland et al. 2006; Koulouriotis and Xanthopoulos 2008), the non-stationary stochastic MAB were investigated under some special settings. For general non-stationary stochastic MAB, Garivier and Moulines (2011) proved that the Discounted UCB algorithm introduced by Kocsis and Szepesvári (2006) attains an $O(K\sqrt{LT}\log T)$ dynamic regret. They also proposed a new algorithm called Sliding Window UCB, for which they derived a slightly better regret bound of $O(K\sqrt{LT\log T})$. A further improved bound of $O(\sqrt{KLT\log(KT)})$ was achieved by an successive elimination method with randomized resets (Allesiardo, Féraud, and Maillard 2017). This bound is also attainable for change-detection based algorithms (Liu, Lee, and Shroff 2018; Cao et al. 2019). While these bounds are optimal in terms of L and T up to logarithmic factors, they hold only when the algorithms are tuned with the number of reward distribution changes L. Recently, this issue was solved by two papers (Auer, Gajane, and Ortner 2019; Chen et al. 2019), which developed algorithms that can achieve optimal regret bounds without prior knowledge of L. A common feature of the above work except for Chen et al. (2019) is that the derived regret bounds are in terms of the number of distribution changes. In the literature, there also exists another line of research (Besbes, Gur, and Zeevi 2014; Karnin and Anava 2016) that focuses on bounding regret with respect to the total variation of reward distributions.

Departing from multi-armed bandits, several work investigates non-stationary stochastic bandits with other formulations including bandits with queries (Yu and Mannor 2009), unimodal bandits (Combes and Proutiere 2014), contextual bandits (Luo et al. 2018; Chen et al. 2019), linear bandits (Cheung, Simchi-Levi, and Zhu 2019; Russac, Vernade, and Cappé 2019; Kim and Tewari 2019; Zhao et al. 2020b), combinatorial bandits (Zhou et al. 2020), and convex bandits (Zhao et al. 2020a). Among them, the work of Yu and Mannor (2009) is closely related to this paper in the sense that they also consider side observations on rewards of unselected arms. The difference is that in their work, under a total query budget the learner can actively query some unselected arms to observe the corresponding rewards, while in this paper, whether an unselected arm's reward is observable is determined by the feedback graph rather than the learner.

Finally, there exists a large body of work on adversarial bandits with graph feedback (Mannor and Shamir 2011; Kocák et al. 2014; Neu 2015; Alon et al. 2017), where the reward of each arm is determined by an adversary and can be thus almost arbitrary. While this reward model is more general, these work define regret with respect to a fixed arm, which is different from the regret used in our non-stationary model that compares the learner against a dynamic sequence of arms.

Preliminary

We study stochastic bandits with graph feedback, where a learner interacts with K arms $\{1, \ldots, K\}$. For an arm $a \in [K]$,[3] we denote by $\mathcal{N}_a$ the union of a and its neighbors arms, i.e., $\mathcal{N}_a = \{a\} \cup \{b \in [K] \colon (a, b) \in E\}$, where E is the edge set of the undirected feedback graph G. The learning protocol proceeds over T rounds. In each round $t \in [T]$, the learner first selects an arm a_t. Then, the learner receives a reward of the chosen arm $r_t(a_t)$ and additionally observes rewards of its neighbors $\{r_t(a) \colon a \in \mathcal{N}_{a_t}, a \neq a_t\}$. For

[2] We use $\tilde{O}(\cdot)$ to hide logarithmic factors.

[3] We use the common notation $[n] = \{1, \ldots, n\}$ for $n \in \mathbb{N}$.

each arm $a \in \mathcal{N}_{a_t}$, its reward $r_t(a)$ is drawn from a distribution $\mathcal{D}_t(a)$ with mean $\mu_t(a)$, i.e., $\mathbb{E}[r_t(a)] = \mu_t(a)$. We assume the rewards are all bounded in $[0, 1]$ and the reward distributions are independent across arms and rounds. Let $\mathcal{A}_t^* = \{a \in [K] \colon \mu_t(a) = \max_{a' \in [K]} \mu_t(a')\}$ and $a_t^* \in \mathcal{A}_t^*$ denote the set consisting of all optimal arms and an optimal arm in round t, respectively. The learner's performance is evaluated by the dynamic regret:

$$\mathrm{DR}(T) = \sum_{t=1}^{T} \big(\mu_t(a_t^*) - \mu_t(a_t)\big).$$

The hardness of the problem is affected by both the non-stationarity of reward distributions and the structure of the feedback graph. The former is captured by L, the number of distribution changes:

$$L = |\{t \in [T] \colon \exists a \in [K], \mathcal{D}_t(a) \neq \mathcal{D}_{t-1}(a)\}|$$

where for notational convenience we define $\mathcal{D}_0(a)$ to be any distribution that is different from $\mathcal{D}_1(a)$ for $a \in [K]$. The latter is characterized by two alternative quantities: the clique covering number θ and the independence number α, which satisfy $\alpha \leq \theta$ and are defined as follows (West et al. 2001).

Definition 1 (Clique Covering Number)**.** *A clique C in a graph $G = (V, E)$ is a subset of V such that every two distinct vertices in C are adjacent. A clique covering in G is a set of cliques $\{C_1, \ldots, C_n\}$ such that $\forall i, j \in [n], C_i \cap C_j = \varnothing$ and $\cup_{i=1}^{n} C_i = V$. The clique covering number θ is defined as the minimum cardinality of a clique covering in G.*

Definition 2 (Independence Number)**.** *An independent set I in a graph $G = (V, E)$ is a subset of V such that for any two distinct vertices in I, there is no edge between them. The independence number α is defined as the maximum cardinality of an independent set in G.*

Warm-up: UCB-NEW

As a warm up, we present a simple extension of the UCB-NE algorithm (Hu, Mehta, and Pan 2019). We first review UCB-NE: In each round t, following the principle of "optimism in the face of uncertainty", UCB-NE selects the arm with the highest sum of empirical mean reward and a confidence term (with ties broken arbitrarily):

$$a_t = \operatorname*{arg\,max}_{a \in [K]} \ \hat{\mu}_{t-1}(a) + c_{t-1}(a).$$

Here, $\hat{\mu}_{t-1}(a)$ is the empirical mean reward of arm a over the first $t-1$ rounds[4]

$$\hat{\mu}_{t-1}(a) = \frac{\sum_{s=1}^{t-1} r_s(a) \mathbb{1}\{a \in \mathcal{N}_{a_s}\}}{O_{t-1}(a)} = \frac{\sum_{s=1}^{t-1} r_s(a) \mathbb{1}\{a_s \in \mathcal{N}_a\}}{O_{t-1}(a)}$$

with $O_{t-1}(a)$ denoting the number of times that the reward of arm a is observed up to round $t-1$

$$O_{t-1}(a) = \sum_{s=1}^{t-1} \mathbb{1}\{a_s \in \mathcal{N}_a\}$$

[4]We use the convention $x/0 = +\infty$ for $x \geq 0$ and denote by $\mathbb{1}\{\cdot\}$ an indicator random variable associated with event $\{\cdot\}$.

Algorithm 1 UCB-NEW

Input: time horizon T, window length ρ
1: Set $\hat{\mu}_0(\rho, a) = 0$ and $c_0(\rho, a) = +\infty$ for each $a \in [K]$
2: **for** $t = 1, \ldots, T$ **do**
3: Play $a_t = \arg\max_{a \in [K]} \ \hat{\mu}_{t-1}(\rho, a) + c_{t-1}(\rho, a)$
4: **end for**

and $c_{t-1}(a)$ is the confidence term defined as

$$c_{t-1}(a) = \sqrt{\frac{2 \log\big(|\mathcal{N}_a|^{1/4}(t-1)\big)}{O_{t-1}(a)}}.$$

While for stationary reward distributions, UCB-NE enjoys sublinear regret bounds, it fails to achieve meaningful dynamic regret bounds in non-stationary environments, since the reward distributions change with time and the mean estimator $\hat{\mu}_{t-1}(a)$ can be far away from the true mean $\mu_t(a)$ for large t. A simple and elegant solution to this issue is the sliding-window mean estimator, which was first introduced by Garivier and Moulines (2011) for multi-armed bandits and has been applied to other bandits problems (Combes and Proutiere 2014; Cheung, Simchi-Levi, and Zhu 2019). Its main idea is to use only the most recent ρ observations when computing the empirical mean reward. We apply it to UCB-NE and replace $\hat{\mu}_{t-1}(a)$ with

$$\hat{\mu}_{t-1}(\rho, a) = \frac{1}{O_{t-1}(\rho, a)} \sum_{s=t-\rho \vee 1}^{t-1} r_s(a) \mathbb{1}\{a_s \in \mathcal{N}_a\}$$

where $t - \rho \vee 1 = \max(t - \rho, 1)$, and $O_{t-1}(\rho, a)$ denotes the number of times that the reward of arm a is observed during the sliding window interval $[t - \rho \vee 1, t-1]$:

$$O_{t-1}(\rho, a) = \sum_{s=t-\rho \vee 1}^{t-1} \mathbb{1}\{a_s \in \mathcal{N}_a\}.$$

The confidence term is correspondingly modified to

$$c_{t-1}(\rho, a) = \sqrt{\frac{3 \log\big(|\mathcal{N}_a|^{1/3}(t - 1 \wedge \rho)\big)}{2 O_{t-1}(\rho, a)}}$$

with $t - 1 \wedge \rho = \min(t-1, \rho)$. Here, we also change the order of $|\mathcal{N}_a|$ and the constant factor, the reason for which will become clear in the theoretical analysis at Appendix B of the full version.[5]

The above procedure is summarized in Algorithm 1, which is named as UCB-NE with sliding Window (UCB-NEW) and has the following theoretical guarantee.

Theorem 1. *The dynamic regret of UCB-NEW satisfies*

$$\mathbb{E}[\mathrm{DR}(T)] \leq \frac{9\theta \log\big(\rho \cdot d_{\max}^{1/3} + 2\big)}{\Delta_{\min}^2} \cdot \left(\frac{T}{\rho} + L\rho + 1\right) + 1$$

where $d_{\max} = \max_{a \in [K]} |\mathcal{N}_a|$ is the maximum degree plus 1 *and $\Delta_{\min} = \min_{t \in [T], a \notin \mathcal{A}_t^*} \mu_t(a_t^*) - \mu_t(a)$ is the minimum*

[5]https://www.lamda.nju.edu.cn/lusy/ns-graph-bandits.pdf

reward gap. Furthermore, when the number of reward distribution changes L is known in advance,[6] *by setting ρ optimally as $\rho = \lceil \sqrt{T/L} \rceil$, UCB-NEW achieves an $\tilde{O}\left(\theta\sqrt{LT}\right)$ dynamic regret bound.*

Improved Algorithm: SEASIDE

While UCB-NEW is simple, its dynamic regret bound is sub-optimal with respect to θ. In this section, we propose an improved algorithm that attains an $\tilde{O}(\sqrt{\alpha LT})$ dynamic regret, which matches the $\Omega\big(\sqrt{\alpha LT}\big)$ lower bound (Mannor and Shamir 2011; Garivier and Moulines 2011; Zhou et al. 2020), up to logarithmic factors. Different from UCB-NEW, the proposed algorithm is built upon the Successive Elimination (SE) framework (Even-Dar, Mannor, and Mansour 2006; Allesiardo, Féraud, and Maillard 2017).

In SE, rounds are divided into epochs:

$$[1, T] = [e_1, e_2) \cup [e_2, e_3) \cup \cdots \cup [e_m, e_{m+1})$$

where $e_\tau, \tau \in [m]$ denotes the beginning of the τ-th epoch and e_{m+1} is defined to be $T + 1$. The basic idea of SE is to maintain an epoch-variant subset of arms $\mathcal{A}_\tau$ and only play arms in $\mathcal{A}_\tau$ during epoch τ. $\mathcal{A}_\tau$ is initialized to be the arm set $[K]$ and gradually shrinks to contain only optimal arms. Specifically, in the τ-th epoch, all arms in $\mathcal{A}_\tau$ are firstly played once to update their empirical mean rewards $\hat{\mu}_\tau(a), a \in \mathcal{A}_\tau$. Let $\tilde{a}^*_\tau$ be an arm with the highest empirical mean reward: $\tilde{a}^*_\tau \in \arg\max_{a \in \mathcal{A}_\tau} \hat{\mu}_\tau(a)$. Then, only arms that are statistically indistinguishable from $\tilde{a}^*_\tau$ are preserved and the other arms are eliminated from $\mathcal{A}_\tau$:

$$\mathcal{A}_{\tau+1} = \left\{ a \in \mathcal{A}_\tau \colon \hat{\mu}_\tau(a) > \hat{\mu}_\tau(\tilde{a}^*_\tau) - 2\sqrt{\frac{\log(KT\tau)}{\tau}} \right\}.$$

In stationary environments where reward distributions remain fixed, it can be shown that with probability $1 - 1/T$, after $O(\log T)$ epochs, all sub-optimal arms are eliminated. As in each epoch, every arm is only played once, the length of each epoch is upper bounded by K. This implies that after $O(K \log T)$ rounds, only optimal arms can be played. Thus, the expected regret can be bounded as $O\big((1 - 1/T) \cdot K \log T + 1/T \cdot T\big) = O(K \log T)$, which is optimal for multi-armed bandits. However, when coming to non-stationary environments, the above analysis becomes invalid. The reason is that in non-stationary environments, the reward distribution of each arm varies with time and thus an eliminated arm can become uniquely optimal at some time, causing linear regrets. We address this problem by using randomized resets (Allesiardo, Féraud, and Maillard 2017): In the end of each round, with a proper probability p, reset SE. In this way, the unique optimal arm that is not in $\mathcal{A}_\tau$ has the chance of returning to $\mathcal{A}_\tau$ and being played.

On the other hand, while under bandit feedback it is necessary to play each arm in A_τ once in order to observe rewards of arms in A_τ, it is inefficient for graph feedback. To mitigate this inefficiency, we employ the AlphaSample strategy (Cohen, Hazan, and Koren 2016). Let S be the set of arms whose rewards need to be observed. AlphaSample repeats the following three steps until S is empty: choosing an arm from S uniformly at random to play, collecting observations of rewards for the chosen arm and its neighbors in S, and removing the chosen arm and its neighbors from S. As the feedback graph is undirected, the arms chosen by AlphaSample constitute an independence set of G. Thus, the number of rounds before AlphaSample terminates is not more than independence number α, which implies an upper bound of α on the length of each epoch, improving the aforementioned bound of K.

We combine randomized resets with AlphaSample to yield Algorithm 2. We termed it as Successive Elimination with AlphaSample and ranDomized rEsets (SEASIDE) and prove an optimal dynamic regret bound for it.

Algorithm 2 SEASIDE

Input: time horizon T, reset probability p

Initialize $t = 1, \tau = 1, e_1 = 1, \tilde{\tau} = 0, \mathcal{A}_1 = [K]$
Set $\hat{\mu}_{\tilde{\tau}}(a) = 0$ for each arm $a \in [K]$
while $t \leq T$ **do**
  Set $S = \mathcal{A}_\tau$
  repeat
    Choose an arm $a \in S$ uniformly at random to play
    Collect observations $o_\tau(a') = r_t(a'), a' \in \mathcal{N}_a \cap S$
    Set $S = S - \mathcal{N}_a$ and $t = t + 1$
    with probability p **do**
      $\tilde{\tau} = \tau, \tau = \tau + 1, e_\tau = t, \mathcal{A}_\tau = [K]$
      Goto Step 2
    end with probability
  until S is empty **or** $t > T$
  if $t > T$ **then**
    Terminate
  end if
  Compute empirical mean rewards as

$$\hat{\mu}_\tau(a) = \frac{(\tau - \tilde{\tau} - 1) \cdot \hat{\mu}_{\tau-1}(a) + o_\tau(a)}{\tau - \tilde{\tau}}, \forall a \in \mathcal{A}_\tau$$

  Find $\tilde{a}^*_\tau \in \arg\max_{a \in \mathcal{A}_\tau} \hat{\mu}_\tau(a)$ and update $\mathcal{A}_{\tau+1} =$

$$\left\{ a \in \mathcal{A}_\tau \colon \hat{\mu}_\tau(a) > \hat{\mu}_\tau(\tilde{a}^*_\tau) - 2\sqrt{\frac{\log(KT(\tau - \tilde{\tau}))}{\tau - \tilde{\tau}}} \right\}$$

  Set $\tau = \tau + 1$ and $e_\tau = t$
end while

Theorem 2. *The dynamic regret of SEASIDE satisfies*

$$\mathbb{E}[\mathrm{DR}(T)] \leq \left(\frac{10}{\Delta^2_{\min}} \log \frac{7KT}{\Delta_{\min}} + \frac{1}{2} \right) \left(\frac{L}{4p} + 4\alpha p T \right) + 2.$$

Furthermore, when the number of reward distribution changes L is known in advance,[7] *by setting p optimally as $p = \sqrt{L/(16\alpha T)}$, SEASIDE achieves an $\tilde{O}(\sqrt{\alpha LT})$ dynamic regret bound.*

[6] Otherwise, we can set $\rho = \lceil \sqrt{T} \rceil$ and obtain a dynamic regret bound of $\tilde{O}(\theta L\sqrt{T})$, which is still sublinear in T.

[7] Otherwise, we can set $p = \sqrt{1/(16\alpha T)}$ and obtain a dynamic regret bound of $\tilde{O}(L\sqrt{\alpha T})$, which is still sublinear in T.

Algorithm 3 ASG

Input: time horizon T

$t = 1, \tau = 1$
$e_\tau = t, \mathcal{G}_t = [K], \mathcal{B}_t = \mathcal{W}_t = \varnothing, \mathcal{S}_t(a) = \varnothing, \forall a \in [K]$
while $t \leq T$ **do** ▷ In epoch τ
  for each $a \in \mathcal{W}_t$ **do** ▷ Add sampling obligations
    for each $\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+$ **do**
      with probability $\epsilon\sqrt{\tau / \left(|\mathcal{W}_t| T \log\left(KT\right)\right)}$ **do**
        $n_\epsilon = \lceil 1.5\epsilon^{-2} \log\left(KT\right) \rceil$
        $S_t(a) = S_t(a) \cup \{(\epsilon, n_\epsilon, t)\}$
      end with probability
    end for
  end for
  Play $a_t = \arg\min_{a \in \mathcal{E}_t} \ \zeta_t(a)$
  Observe rewards of arms in $\mathcal{N}_{a_t}$
  for each $a \in \mathcal{B}_t$ **do** ▷ Update $\mathcal{S}_t(a)$
    $S_{t+1}(a) = \{(\epsilon, n_\epsilon, s) \in \mathcal{S}_t(a) \colon n_{[s,t]}(a) < n_\epsilon\}$
  end for
  for each $a \in \mathcal{G}_t$ **do** ▷ Detect changes for good arms
    if there is $s_1, s_2, s \in [e_\tau, t]$ such that (5) holds **then**
      $t = t + 1, \tau = \tau + 1$, goto Step 2
    end if
  end for
  for each $a \in \mathcal{B}_t$ **do** ▷ Detect changes for bad arms
    if there is $s \in [e_\tau, t]$ such that (4) holds **then**
      $t = t + 1, \tau = \tau + 1$, goto Step 2
    end if
  end for
  for each $a \in \mathcal{G}_t$ **do** ▷ Eliminate some good arms
    if there is $s \in [e_\tau, t]$ such that (3) holds **then**
      $\mathcal{B}_t = \mathcal{B}_t \cup \{a\}$, store $\tilde{\mu}_\tau(a)$ and $\tilde{\Delta}_\tau(a)$
    end if
  end for
  $\mathcal{B}_{t+1} = \mathcal{B}_t, \mathcal{G}_{t+1} = [K] - \mathcal{B}_{t+1}$
  $\mathcal{W}_{t+1} = \text{ComputeW}(\mathcal{B}_{t+1}, \mathcal{G}_{t+1}), t = t + 1$
end while

Remark 1. While in this paper we assume undirected feedback graphs, by leveraging the analysis of AlphaSample (Cohen, Hazan, and Koren 2016), we can derive a high probability bound on the length of each epoch under the more general setting where the feedback graph is directed. Based on this bound, we will prove a variant of Theorem 2 for general directed feedback graphs at Appendix C.

Adaptive Algorithm: ASG

To achieve the $\tilde{O}\big(\sqrt{\alpha LT}\big)$ dynamic regret bound, SEASIDE needs to know the number of reward distribution changes L in advance. Without such prior knowledge, the regret bound of SEASIDE will scale with L instead of $\sqrt{L}$. In this section, we develop an adaptive algorithm called ASG that can achieve an $\tilde{O}(\sqrt{\theta LT})$ dynamic regret bound without prior knowledge of L. ASG follows the algorithmic framework of Auer, Gajane, and Ortner (2019), but with a novel sampling scheme and a corresponding arm selection strategy that can exploit the graph feedback.

Algorithm 4 ComputeW

Input: bar arm set $\mathcal{B}_{t+1}$, good arm set $\mathcal{G}_{t+1}$, epoch index τ

$\tilde{\mathcal{B}} = \mathcal{B}_{t+1}, \tilde{\mathcal{W}} = \varnothing$
while $\tilde{\mathcal{B}} \neq \varnothing$ **do**
  Find $a \in \arg\min_{a' \in \tilde{\mathcal{B}}} \tilde{\Delta}_\tau(a')$
  Update $\tilde{\mathcal{W}} = \tilde{\mathcal{W}} \cup \{a\}, \ \tilde{\mathcal{B}} = \tilde{\mathcal{B}} - \mathcal{N}_a$
end while
return $\tilde{\mathcal{W}}$

Before presenting ASG, we introduce some definitions. Let $O_{[s,t]}(a)$ be the number of times that the reward of arm a is observed during $[s, t]$: $O_{[s,t]}(a) = \sum_{i=s}^{t} \mathbb{1}\{a_i \in \mathcal{N}(a)\}$. We denote the empirical mean reward of a over $[s, t]$ by

$$\hat{\mu}_{[s,t]}(a) = \frac{\sum_{i=s}^{t} r_i(a) \mathbb{1}\{a_i \in \mathcal{N}(a)\}}{O_{[s,t]}(a)}.$$

With a slight abuse of notation, for a set of arms $A \subseteq [K]$, we define $O_{[s,t]}(A)$ as the maximum o such that for any arm in A, its reward is observed at least o times during $[s, t]$

$$O_{[s,t]}(A) = \max\{o \in \mathbb{N} \colon \forall a \in A, O_{[s,t]}(a) \geq o\}. \quad (1)$$

An equivalent definition is $O_{[s,t]}(A) = \min_{a \in A} O_{[s,t]}(a)$. Finally, we denote by $n_{[s,t]}(a)$ the number of times that arm a is played during $[s, t]$.

We are now ready to present ASG, which is outlined in Algorithm 3. To handle non-stationary reward distributions with unknown number of changes, ASG performs change detection test for reward distributions in each round and restarts once it detects a change. Let $e_1 < \cdots < e_m$ denote the rounds when ASG restarts, i.e., Step 2 is executed. We can divide $[1, T]$ into epochs as follows

$$[1, T] = [e_1, e_2) \cup [e_2, e_3) \cup \cdots \cup [e_m, e_{m+1}) \quad (2)$$

where we define $e_{m+1} = T + 1$ and $[e_\tau, e_{\tau+1})$ is the τ-th epoch. In each epoch, ASG splits the arm set $[K]$ into two time-variant subsets: a good arm set $\mathcal{G}_t$ and a bad arm set $\mathcal{B}_t$. In the beginning of epoch τ, all arms are good, i.e., $\mathcal{G}_{e_\tau} = [K], \ \mathcal{B}_{e_\tau} = \varnothing$. During epoch τ, as time goes, arms whose empirical mean rewards are significantly worse than that of the seemingly optimal arm are eliminated from the good arm set and added into the bad arm set. More precisely, in round $t \in [e_\tau, e_{\tau+1} - 1]$, for an arm $a \in \mathcal{G}_t$, if there is $s \in [e_\tau, t]$ such that

$$\max_{a' \in \mathcal{G}_t} \hat{\mu}_{[s,t]}(a') - \hat{\mu}_{[s,t]}(a) > \left(\sqrt{2} + 1\right) \sqrt{\frac{6 \log\left(KT\right)}{O_{[s,t]}(\mathcal{G}_t)}}. \quad (3)$$

Then, it is removed from $\mathcal{G}_t$ and added into $\mathcal{B}_t$. Furthermore, its empirical mean reward and the gap to the seemingly optimal arm are stored as $\tilde{\mu}_\tau(a) = \hat{\mu}_{[s,t]}(a)$ and $\tilde{\Delta}_\tau(a) = \max_{a' \in \mathcal{G}_t} \hat{\mu}_{[s,t]}(a') - \hat{\mu}_{[s,t]}(a)$, which will be used in the subsequent rounds for detecting changes of its reward distribution.

Specifically, in each round $t \in [e_\tau, e_{\tau+1} - 1]$, for every bad arm $a \in \mathcal{B}_t$, ASG performs the following change detection test: whether there is $s \in [e_\tau, t]$ such that the inequality

$$|\hat{\mu}_{[s,t]}(a) - \tilde{\mu}_\tau(a)| > \frac{\tilde{\Delta}_\tau(a)}{4} + \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}} \quad (4)$$

holds.[8] If yes, ASG concludes that the reward distribution of a has changed and consequently enters into a new epoch where everything is reset. On the other hand, it is also necessary to detect changes for good arms. For a good arm $a \in \mathcal{G}_t$, its reward distribution is detected to have changed if there is $s_1, s_2, s \in [e_\tau, t]$ such that the following inequality holds

$$|\hat{\mu}_{[s_1,s_2]}(a) - \hat{\mu}_{[s,t]}(a)| > \sqrt{\frac{3\log(KT)}{2O_{[s_1,s_2]}(\mathcal{G}_{s_2})}} + \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(\mathcal{G}_t)}}. \quad (5)$$

Intuitively, for an arm a, to quickly detect the change of its reward distribution, we should collect many recent observations of its reward by playing arms in $\mathcal{N}(a)$ often. However, if arms in $\mathcal{N}(a)$ are all bad arms, i.e., $\mathcal{N}(a) \subseteq \mathcal{B}_t$, and the reward distributions do not change, doing so may cause large regret. A solution to this dilemma is the consecutive sampling policy proposed by Auer, Gajane, and Ortner (2019). The main idea is to maintain a time-variant set $\mathcal{S}_t(a)$ for each bad arm $a \in \mathcal{B}_t$, and in round t choose the arm from $\{a\colon a \in \mathcal{G}_t \text{ OR } \mathcal{S}_t(a) \neq \varnothing\}$ in a round-robin fashion. Each item in $\mathcal{S}_t(a)$ is a triple $(\epsilon, n_\epsilon, s)$ called sampling obligation, where $\epsilon \in \{1/2, 1/4, 1/8, \ldots\}$ is the magnitude of reward distribution change that we aim to detect for arm a, $n_\epsilon = \lceil 1.5\epsilon^{-2}\log(KT)\rceil$ is the number of samples required to detect such change, and s is the time that the sampling obligation is added into $\mathcal{S}_t(a)$.

The set $\mathcal{S}_t(a)$ is initialized to be empty and is updated in each round after a becomes bad. Specifically, in round t when $a \in \mathcal{B}_t$, for every $\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+$, with probability $\epsilon\sqrt{\tau/(KT\log(KT))}$, we add $(\epsilon, n_\epsilon, t)$ into $\mathcal{S}_t(a)$. For a bad arm a with sampling obligation $(\epsilon, n_\epsilon, s)$, if a has been played n times during $[s, t]$, we remove $(\epsilon, n_\epsilon, s)$ from $\mathcal{S}_t(a)$. The advantage of the above policy, as stated by Auer, Gajane, and Ortner (2019), is as follows. First, when the reward distribution of bad arm a has changed, if the change is small, it causes small regret. If the change is large, it will be detected in time and the expected regret can be bounded, as the probability of adding sampling obligations scales linearly with ϵ. Second, when the reward distribution of a does not change, the expected regret caused by playing a can be also controlled, since n_ϵ is on the order of $1/\epsilon^2$.

While in multi-armed bandits setting, the policy of Auer, Gajane, and Ortner (2019) leads to an optimal $\tilde{O}(\sqrt{KLT})$ dynamic regret bound without knowing L, to achieve an improved $\tilde{O}(\sqrt{\theta LT})$ bound for the setting considered in this paper, it has to be extended to exploit graph feedback.

[8] We use the convention that for $B = +\infty$, $A > B$ is false and $A \leq B$ is true regardless the value of A.

We here propose a non-trivial extension of Auer, Gajane, and Ortner (2019) that can utilize graph feedback. The basic idea is to maintain a subset $\mathcal{W}_t$ of bad arms and to apply consecutive sampling on this subset: In each round t, we only add sampling obligations for arms in $\mathcal{W}_t$. The specific mechanism of adding sampling obligations into $\mathcal{S}_t(a)$ for $a \in \mathcal{W}_t$ is similar to that in the aforementioned consecutive sampling policy, with the main difference of modifying the probability from $\epsilon\sqrt{\tau/(KT\log(KT))}$ to $\epsilon\sqrt{\tau/(|\mathcal{W}_t|T\log(KT))}$.

The set $\mathcal{W}_t$ is initialized to be empty at the beginning of an epoch. At each time when new arms are added into the bad arm set, i.e., $\mathcal{B}_{t+1} \neq \mathcal{B}_t$, we update $\mathcal{W}_t$ by invoking Algorithm 4, which proceeds as follows. First, Algorithm 4 creates two auxiliary sets $\tilde{\mathcal{B}}, \tilde{\mathcal{W}}$ and sets $\tilde{\mathcal{B}} = \mathcal{B}_{t+1}, \tilde{\mathcal{W}} = \varnothing$. Then, the algorithm repeats the following three steps until $\tilde{\mathcal{B}}$ is empty: choosing an arm a from $\mathcal{B}$ with the minimum gap $\tilde{\Delta}_\tau(a)$, adding a into $\tilde{\mathcal{W}}$, and removing a as well as its neighbors from $\tilde{\mathcal{B}}$. Finally, Algorithm 4 returns $\tilde{\mathcal{W}}$, which is used to set $\mathcal{W}_{t+1} = \tilde{\mathcal{W}}$. The intuition behind the design of Algorithm 4 is two folds. First, Algorithm 4 ensures that any two arms in $\mathcal{W}_{t+1}$ are not adjacent. So the size of $\mathcal{W}_{t+1}$ never exceeds the independence number α. Second, for any bad arm $a \in \mathcal{B}_{t+1} - \mathcal{W}_{t+1}$, by Algorithm 4, there must be an arm $a' \in \mathcal{N}_a \cap \mathcal{W}_{t+1}$ with $\tilde{\Delta}_\tau(a') \leq \tilde{\Delta}_\tau(a)$. Thus, for every magnitude $\epsilon \geq \tilde{\Delta}_\tau(a)/16$ of reward distribution change that we aim to detect for a, it holds that $\epsilon \geq \tilde{\Delta}_\tau(a')/16$, which implies that for $s < t+1$ the sampling obligation $(\epsilon, n_\epsilon, s)$ is added into $\mathcal{S}_s(a')$ with some probability. From the perspective of collecting reward observations for a, adding $(\epsilon, n_\epsilon, s)$ into $\mathcal{S}_s(a')$ can be viewed as adding $(\epsilon, n_\epsilon, s)$ into $\mathcal{S}_s(a)$, since reward of a can be observed by playing a'.

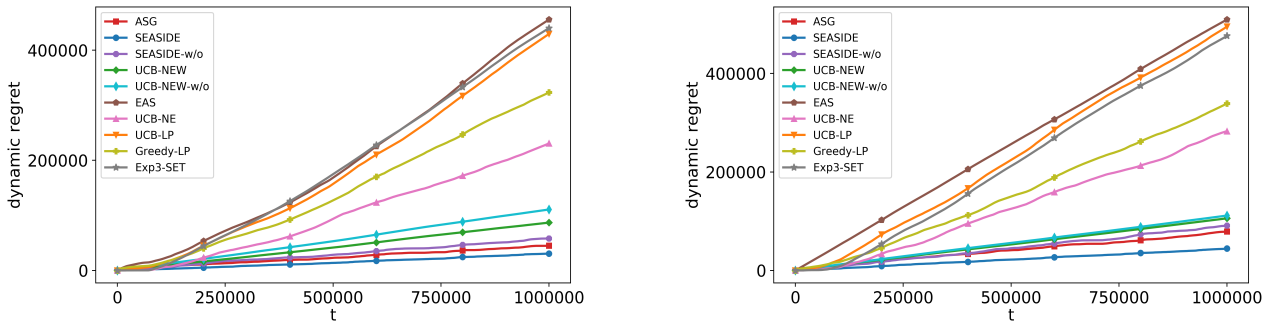
It remains to describe the arm selection strategy. In each round t, only good arms and bad arms with nonempty $\mathcal{S}_t(a)$ can be played. We denote by $\mathcal{E}_t$ the set comprised of these arms: $\mathcal{E}_t = \{a \in [K]\colon a \in \mathcal{G}_t \text{ OR } \mathcal{S}_t(a) \neq \varnothing\}$. Following Auer, Gajane, and Ortner (2019), we call arms in $\mathcal{E}_t$ as eligible arms. For an eligible arm $a \in \mathcal{E}_t$, let $\zeta_t(a)$ be the last time when reward of a is observed: $\zeta_t(a) = \min\{s \in \mathbb{N}\colon \mathcal{N}_a \cap \{a_{s+1}, a_{s+2}, \ldots, a_{t-1}\} = \varnothing\}$. ASG plays the eligible arm with the minimum $\zeta_t(a)$: $a_t = \arg\min_{a \in \mathcal{E}_t} \zeta_t(a)$, where ties are broken by giving priority to arms with sampling obligations. In other words, a_t is the eligible arm whose reward is observed least recently. The advantages of this arm selection rule are summarized in Lemmas 4 and 5 at Appendix E, which play a key role in the regret analysis.

Finally, we present the theoretical guarantee of ASG.

Theorem 3. *Without knowing the number of reward distribution changes L, ASG achieves the following dynamic regret bound*

$$\mathbb{E}[\mathrm{DR}(T)] \leq \tilde{O}\big(\sqrt{\theta LT}\big).$$

Remark 2. Following the suggestion in Auer, Gajane, and Ortner (2019), in practice we can reduce the time complexity of ASG to $O(K\log^2 T)$ per step by checking (3), (4), and (5) for only intervals of certain length $(2^h, h = 1, \ldots, \lfloor\log_2 T\rfloor)$. To check (3) and (4) for each length 2^h, we only need to compute, for each arm a, the cumulative



(a) $L = 10$ (b) $L = 20$

Figure 1: Dynamic regret of the examined algorithms

observed rewards of a during $[t - 2^h + 1, t]$ and the total times that the reward of a is observed during $[t - 2^h + 1, t]$. This computation can be performed in an online manner with $O(1)$ time complexity by maintaining two queues of length 2^h: one for storing the observed reward $r_i(a)$ and the other for storing the indicator variable $\mathbb{1}\{a_i \in \mathcal{N}_a\}$. To check (5) for each interval length 2^h, we only need to additionally keep $2K$ variables $\hat{\mu}_h^+(a)$ and $\hat{\mu}_h^-(a)$ for $a = 1, \ldots, K$ that store the minimum value of $\hat{\mu}_{[s-2^h+1,s]}(a) + \sqrt{3\log(KT)/(2O_{[s-2^h+1,s]}(\mathcal{G}_s))}$ and the maximum value of $\hat{\mu}_{[s-2^h+1,s]}(a) - \sqrt{3\log(KT)/(2O_{[s-2^h+1,s]}(\mathcal{G}_s))}$ over all intervals of length 2^h, respectively. These $2K$ variables can be also updated together with $O(K)$ time complexity in each round t based on the cumulative observed rewards during $[t - 2^h + 1, t]$ of each arm and the total times that the reward is observed during $[t - 2^h + 1, t]$ of each arm.

Experiment

In this section, we present experimental results to illustrate the empirical performance of our proposed algorithms. For UCB-NEW and SEASIDE which require prior knowledge of L to achieve the regret bounds scaling with $\sqrt{L}$, we examine two versions, i.e., one with and the other without tuning parameters in terms of L. For ASG, we reduce its computational cost according to Remark 2. We adopt UCB-NE (Hu, Mehta, and Pan 2019), Exp3-SET (Alon et al. 2017), EAS (Elimination with AlphaSample, Cohen, Hazan, and Koren 2016), Greedy-LP and UCB-LP (Buccapatnam, Eryilmaz, and Shroff 2014) as baseline algorithms.

We use a synthetic dataset constructed as follows. Let $K = 30$, $T = 1000000$ and pick L from $\{10, 20\}$. We first randomly choose $L - 1$ breakpoints from $\{2, \ldots, T - 1\}$ to partition $[1, T]$ into L stationary intervals. Then, in each stationary interval, we choose 2 arms uniformly at random from $[K]$ as optimal arms and set their mean rewards to be 0.9. For suboptimal arms, their mean rewards are sampled from a uniform distribution with support $[0, 0.7]$. For each arm, we generate its rewards by drawing samples from truncated normal distributions with support $[0, 1]$ and variance 0.01. Finally, inspired by Brigham and Dutton (1983), we construct a feedback graph illustrated at Appendix A with $\alpha = 10$ and $\theta = 14$.

We run each algorithm 10 times and report the average performance in Fig. 1, where "-w/o" stands for "without prior knowledge of L". As can be seen, our proposed algorithms significantly outperform the baseline algorithms, which is expected since the baseline algorithms assume the reward of each arm is either drawn from a stationary distribution or determined by an adversary. Furthermore, without prior knowledge of L, UCB-NEW and SEASIDE still behave well and achieve much smaller regrets than the baseline algorithms, demonstrating their practicality. Finally, while SEASIDE attains the smallest regret with prior knowledge of L, it becomes inferior to ASG when L is unknown, which validates the advantage of ASG's adaptivity.

Conclusion and Future Work

We have presented three algorithms for stochastic bandits with graph feedback in non-stationary environments. The first algorithm is simple but only achieves a sub-optimal regret bound. The second and the third algorithms, though much more complicated, enjoy regret bounds matching the lower bounds and each has its own advantage: The second algorithm enjoys a better regret bound which depends on the independence number α and holds for general directed feedback graphs, but it needs to know the number of reward distribution changes L in advance. By contrast, the third algorithm requires no prior knowledge of L, but its regret bound is in terms of the clique covering number $\theta \geq \alpha$ and only applies to the setting where the feedback graph is undirected.

Thus, a natural and challenging open problem is to design a parameter-free algorithm with $\tilde{O}\big(\sqrt{\alpha LT}\big)$ regret bounds for directed feedback graphs, which we leave as a future work. While we currently assume bounded rewards, in the future we will study unbounded and even heavy-tailed reward distributions (Bubeck, Cesa-Bianchi, and Lugosi 2013; Lu et al. 2019). Finally, it is also worthy of pursing to investigate whether the undesired $\sqrt{\theta}$ factor in the regret bound of UCB policy can be removed and whether UCB policy can be extended to directed feedback graphs, since compared to elimination based algorithms, UCB policy is simpler to understand and easier to implement.

Acknowledgments

This work was partially supported by the NSFC (61976112), JiangsuSF (BK20200064), Open Research Projects of Zhejiang Lab (NO. 2021KB0AB02), and Alibaba Innovative Research Program. We thank the anonymous reviewers for their constructive suggestions.

References

Alami, R. 2019. Non-Stationary Thompson Sampling For Stochastic Bandits with Graph-Structured Feedback. URL https://hal.archives-ouvertes.fr/hal-01987001. Working paper or preprint.

Allesiardo, R.; Féraud, R.; and Maillard, O.-A. 2017. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics* 3(4): 267–283.

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.

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.; Gajane, P.; and Ortner, R. 2019. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the 32nd Conference on Learning Theory*, 138–158.

Besbes, O.; Gur, Y.; and Zeevi, A. 2014. Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. In *Advances in Neural Information Processing Systems 27*, 199–207.

Bnaya, Z.; Puzis, R.; Stern, R.; and Felner, A. 2013. Bandit algorithms for social network queries. In *Proceedings of the 2013 International Conference on Social Computing*, 148–153.

Brigham, R. C.; and Dutton, R. D. 1983. On clique covers and independence numbers of graphs. *Discrete Mathematics* 44(2): 139–144.

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

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.

Cao, Y.; Wen, Z.; Kveton, B.; and Xie, Y. 2019. Nearly Optimal Adaptive Procedure with Change Detection for Piecewise-Stationary Bandit. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 418–427.

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.

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

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 *Proceedings of the 32nd Conference on Learning Theory*, 696–726.

Cheung, W. C.; Simchi-Levi, D.; and Zhu, R. 2019. Learning to Optimize under Non-Stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 1079–1087.

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.

Combes, R.; and Proutiere, A. 2014. Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, 521–529.

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.

Garivier, A.; and Moulines, E. 2011. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, 174–188.

Hartland, C.; Gelly, S.; Baskiotis, N.; Teytaud, O.; and Sebag, M. 2006. Multi-armed bandit, dynamic environments and meta-bandits. In *Advances in Neural Information Processing Systems 19 Workshop, Online Trading between Exploration and Exploitation*.

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*.

Jadbabaie, A.; Rakhlin, A.; Shahrampour, S.; and Sridharan, K. 2015. Online Optimization : Competing with Dynamic Comparators. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 398–406.

Karnin, Z. S.; and Anava, O. 2016. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems 29*, 199–207.

Kim, B.; and Tewari, A. 2019. Near-optimal Oracle-efficient Algorithms for Stationary and Non-Stationary Stochastic Linear Bandits. *arXiv preprint arXiv:1912.05695* .

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.

Kocsis, L.; and Szepesvári, C. 2006. Discounted ucb. In *Proceedings of the 2nd PASCAL Challenges Workshop*.

Kocsis, L.; Szepesvári, C.; and Willemson, J. 2006. Improved monte-carlo search. *Univ. Tartu, Estonia, Tech. Rep* 1.

Koulouriotis, D. E.; and Xanthopoulos, A. 2008. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation* 196(2): 913–922.

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

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th 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 *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 3643–3650.

Liu, F.; Lee, J.; and Shroff, N. 2018. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

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.

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.

Luo, H.; Wei, C.-Y.; Agarwal, A.; and Langford, J. 2018. Efficient Contextual Bandits in Non-stationary Worlds. In *Proceedings of the 31st Conference On Learning Theory*, 1739–1776.

Lykouris, T.; Tardos, E.; and Wali, D. 2019. Graph regret bounds for Thompson Sampling and UCB. *arXiv preprint arXiv:1905.09898* .

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.

Min, S.-H.; and Han, I. 2005. Detection of the customer time-variant pattern for improving recommender systems. *Expert Systems with Applications* 28(2): 189–199.

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.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58(5): 527–535.

Russac, Y.; Vernade, C.; and Cappé, O. 2019. Weighted Linear Bandits for Non-Stationary Environments. In *Advances in Neural Information Processing Systems 32*, 12017–12026.

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*.

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

Yu, J. Y.; and Mannor, S. 2009. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th International Conference on Machine Learning*, 1177–1184.

Zeng, C.; Wang, Q.; Mokhtari, S.; and Li, T. 2016. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2025–2034.

Zhao, P.; Wang, G.; Zhang, L.; and Zhou, Z.-H. 2020a. Bandit Convex Optimization in Non-stationary Environments. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 1508–1518.

Zhao, P.; Zhang, L.; Jiang, Y.; and Zhou, Z.-H. 2020b. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*.

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 *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.