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.

1 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).

2 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.

3 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 $\mathcal{A}_t^* = \{a \in [K]: \mu_t(a) = \max_{a' \in [K]} \mu_t(a')\}$ 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:

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

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.*

4 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.

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.[5]

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]All proofs are postponed to the appendices.

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.*

5 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 $\Omega\left(\sqrt{\alpha LT}\right)$ 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 $\mathcal{A}_\tau$:

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

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.

Algorithm 2 SEASIDE

Input: time horizon T, reset probability p

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

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

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

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

19: Set $\tau = \tau + 1$ and $e_\tau = t$
20: **end while**

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.

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

1: $t = 1, \tau = 1$
2: $e_\tau = t, \mathcal{G}_t = [K], \mathcal{B}_t = \mathcal{W}_t = \varnothing, \mathcal{S}_t(a) = \varnothing, \forall a \in [K]$
3: **while** $t \le T$ **do** ▷ In epoch τ
4: **for each** $a \in \mathcal{W}_t$ **do** ▷ Add sampling obligations
5: **for each** $\epsilon = 2^{-g} \ge \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+$ **do**
6: **with probability** $\epsilon\sqrt{\tau/\left(|\mathcal{W}_t|\, T \log{(KT)}\right)}$ **do**
7: $n_\epsilon = \lceil 1.5\epsilon^{-2} \log{(KT)} \rceil$
8: $S_t(a) = S_t(a) \cup \{(\epsilon, n_\epsilon, t)\}$
9: **end with probability**
10: **end for**
11: **end for**
12: Play $a_t = \arg\min_{a \in \mathcal{E}_t} \zeta_t(a)$
13: Observe rewards of arms in $\mathcal{N}_{a_t}$
14: **for each** $a \in \mathcal{B}_t$ **do** ▷ Update $\mathcal{S}_t(a)$
15: $S_{t+1}(a) = \{(\epsilon, n_\epsilon, s) \in \mathcal{S}_t(a) \colon n_{[s,t]}(a) < n_\epsilon\}$
16: **end for**
17: **for each** $a \in \mathcal{G}_t$ **do** ▷ Detect changes for good arms
18: **if** there is $s_1, s_2, s \in [e_\tau, t]$ such that (5) holds **then**
19: $t = t + 1, \tau = \tau + 1$, goto Step 2
20: **end if**
21: **end for**
22: **for each** $a \in \mathcal{B}_t$ **do** ▷ Detect changes for bad arms
23: **if** there is $s \in [e_\tau, t]$ such that (4) holds **then**
24: $t = t + 1, \tau = \tau + 1$, goto Step 2
25: **end if**
26: **end for**
27: **for each** $a \in \mathcal{G}_t$ **do** ▷ Eliminate some good arms
28: **if** there is $s \in [e_\tau, t]$ such that (3) holds **then**
29: $\mathcal{B}_t = \mathcal{B}_t \cup \{a\}$, store $\tilde{\mu}_\tau(a)$ and $\tilde{\Delta}_\tau(a)$
30: **end if**
31: **end for**
32: $\mathcal{B}_{t+1} = \mathcal{B}_t, \mathcal{G}_{t+1} = [K] - \mathcal{B}_{t+1}$
33: $\mathcal{W}_{t+1} = \text{ComputeW}(\mathcal{B}_{t+1}, \mathcal{G}_{t+1}), t = t + 1$
34: **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.

6 Adaptive Algorithm: ASG

To achieve the $\tilde{O}\left(\sqrt{\alpha L T}\right)$ 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 $\tilde{O}(\sqrt{\theta L T})$ 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

1: $\tilde{\mathcal{B}} = \mathcal{B}_{t+1}, \tilde{\mathcal{W}} = \varnothing$
2: **while** $\tilde{\mathcal{B}} \neq \varnothing$ **do**
3: Find $a \in \arg\min_{a' \in \tilde{\mathcal{B}}} \tilde{\Delta}_\tau(a')$
4: Update $\tilde{\mathcal{W}} = \tilde{\mathcal{W}} \cup \{a\},\ \tilde{\mathcal{B}} = \tilde{\mathcal{B}} - \mathcal{N}_a$
5: **end while**
6: **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 $[s,t]$

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

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

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

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

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

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 $\epsilon \in \{1/2, 1/4, 1/8, \dots\}$ 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'.

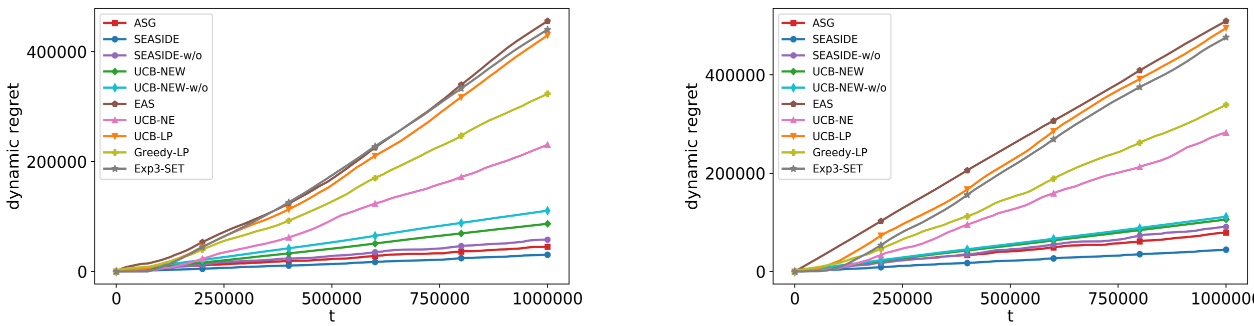
It remains to describe the arm selection strategy. In each round t, only good arms and bad arms with nonempty $\mathcal{S}_t(a)$ can be played. We denote by $\mathcal{E}_t$ the set comprised of these arms: $\mathcal{E}_t = \{a \in [K]\colon a \in \mathcal{G}_t \text{ OR } \mathcal{S}_t(a) \neq \varnothing\}$. 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: $\zeta_t(a) = \min\{s \in \mathbb{N}\colon \mathcal{N}_a \cap \{a_{s+1}, a_{s+2}, \dots, a_{t-1}\} = \varnothing\}$. 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 $(2^h, h = 1, \dots, \lfloor\log_2 T\rfloor)$. 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.

7 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.

8 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*.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* 58(301): 13–30.

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.

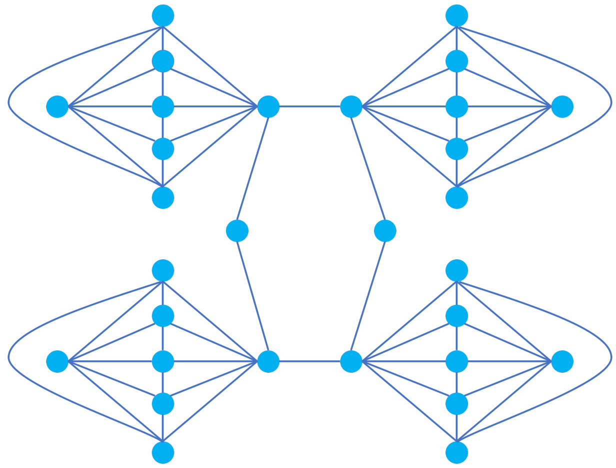
Zeng, C.; Wang, Q.; Mokhtari, S.; and Li, T. 2016. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2025–2034.

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*.

A Illustration of the Feedback Graph Used in the Experiment



B Proof of Theorem 1

While UCB-NEW is a variant of UCB-NE, its proof follows the analysis framework of sliding-window estimator (Garivier and Moulines 2011). The novelty of our analysis is that we partition the arms set into cliques and bound the regret for each clique, which is different from the analysis in Garivier and Moulines (2011), where the regret is bounded separately for each arm. Furthermore, our analysis is easier to follow in the sense that we use standard Hoeffding's inequality (Hoeffding 1963) instead of the complicated concentration inequality proposed by Garivier and Moulines (2011). Finally, we would like to remark that UCB-NE is a variant of UCB-N (Caron et al. 2012), with the main difference being the extra exploration term $\log |\mathcal{N}_a|$. While Hu, Mehta, and Pan (2019) showed that this extra term can remove the linear dependence on K in the *static* regret bound of UCB-N, our analysis reveals that it is also essential in obtaining a *dynamic* regret bound irrespective of K. In fact, if one applies sliding-window to UCB-N rather than UCB-NE, one can only get a worse dynamic regret bound of $\tilde{O}(\theta\sqrt{KLT})$.

Let $\mathcal{C} = \{C_1, C_2, \ldots, C_\theta\}$ be a minimum clique covering of the feedback graph G and recall that $\mathcal{A}_t^*$ denotes the set consisting of all optimal arms in round t

$$\mathcal{A}_t^* = \left\{ a \in [K] \colon \mu_t(a) = \max_{a' \in [K]} \mu_t(a') \right\}.$$

Since rewards are all bounded in $[0, 1]$, we have

$$\begin{aligned} \mathrm{DR}(T) = \sum_{t=1}^{T} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) &= 1 + \sum_{t=2}^{T} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) \\ &\le 1 + \sum_{t=2}^{T} \mathbb{1}\{a_t \notin \mathcal{A}_t^*\} \\ &\le 1 + \sum_{t=2}^{T} \sum_{C \in \mathcal{C}} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*\} = 1 + \sum_{C \in \mathcal{C}} \sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*\}. \end{aligned} \tag{6}$$

Below we bound the term $\sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*\}$ for each clique $C \in \mathcal{C}$.

Fix $C \in \mathcal{C}$. Let $n_{t-1}(\rho, C)$ denote the number of times that an arm in C is played during rounds $[t - \rho \vee 1, t - 1]$

$$n_{t-1}(\rho, C) = \sum_{s = t - \rho \vee 1}^{t-1} \mathbb{1}\{a_s \in C\}$$

and define

$$\Gamma(\rho) = \frac{6 \log \left(\rho \cdot d_{\max}^{1/3} + 2\right)}{\Delta_{\min}^2}$$

where recall that $\Delta_{\min}$ denotes the minimum gap of expected reward between the optimal arm and sub-optimal arms

$$\Delta_{\min} = \min_{t \in [T], a \notin \mathcal{A}_t^*} \mu_t(a_t^*) - \mu_t(a)$$

and $d_{\max}$ denotes the maximum degree of G plus 1

$$d_{\max} = \max_{a \in [K]} |\mathcal{N}_a|.$$

Then, we can split the term $\sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*\}$ as

$$\sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*\} = \underbrace{\sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) < \Gamma(\rho)\}}_{B_1} + \underbrace{\sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \geq \Gamma(\rho)\}}_{B_2}. \tag{7}$$

We first analyze B_1:

$$B_1 \leq \sum_{t=2}^{T} \mathbb{1}\{a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)\} \leq \sum_{i=1}^{\lceil T/\rho \rceil} \sum_{t=(i-1)\rho+1 \vee 2}^{i\rho \wedge T} \mathbb{1}\{a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)\}. \tag{8}$$

For $i \in \{1, \ldots, \lceil T/\rho \rceil\}$, suppose that there exists $t \in [(i-1)\rho+1 \vee 2, i\rho \wedge T]$ such that the event $\{a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)\}$ occurs and define ν_i as the maximum of such t

$$\nu_i = \max\{t \in [(i-1)\rho + 1 \vee 2, i\rho \wedge T] \colon a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)\}.$$

We have

$$\begin{aligned}
\sum_{t=(i-1)\rho+1\vee 2}^{i\rho \wedge T} \mathbb{1}\{a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)\} &= \sum_{t=(i-1)\rho+1\vee 2}^{\nu_i} \mathbb{1}\{a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)\} \\
&\leq \sum_{t=(\nu_i-\rho+1)\vee 2}^{\nu_i} \mathbb{1}\{a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)\} \\
&\leq \sum_{t=(\nu_i-\rho+1)\vee 1}^{\nu_i} \mathbb{1}\{a_t \in C\} \\
&= n_{\nu_i}(\rho, C) \leq n_{\nu_i - 1}(\rho, C) + 1 < \Gamma(\rho) + 1.
\end{aligned}$$

On the other hand, the above inequality also trivially holds for the case that no $t \in [(i-1)\rho + 1 \vee 2, i\rho \wedge T]$ satisfies $a_t \in C, n_{t-1}(\rho, C) < \Gamma(\rho)$. Substituting the above inequality into (8) gives

$$B_1 \leq \sum_{i=1}^{\lceil T/\rho \rceil} \big(\Gamma(\rho) + 1\big) = \left\lceil \frac{T}{\rho} \right\rceil \big(\Gamma(\rho) + 1\big) \leq \left(\frac{T}{\rho} + 1\right) \cdot \left(\frac{6 \log\big(\rho \cdot d_{\max}^{1/3} + 2\big)}{\Delta_{\min}^2} + 1\right) \tag{9}$$

where the last inequality follows from the definition of $\Gamma(\rho)$.

We now turn to bound B_2 and start with introducing $S(\rho)$, which is the set consisting of all rounds $t \in [2, T]$ such that during $[t - \rho \vee 1, t]$, reward distributions of all arms remain fixed

$$S(\rho) = \{t \in [2, T] \colon \forall i, j \in [t - \rho \vee 1, t], \forall a \in [K], \mathcal{D}_i(a) = \mathcal{D}_j(a)\}.$$

We also define $\bar{S}(\rho) = \{t \in [2, T] \colon t \notin S(\rho)\}$ as the complementary set of $S(\rho)$. The reason for introducing $S(\rho)$ and $\bar{S}(\rho)$ is to analyze B_2 for stationary rounds and non-stationary rounds separately:

$$\begin{aligned}
B_2 &= \sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \geq \Gamma(\rho)\} \\
&\leq \sum_{t \in S(\rho)} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \geq \Gamma(\rho)\} + \sum_{t \in \bar{S}(\rho)} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \geq \Gamma(\rho)\}.
\end{aligned} \tag{10}$$

Recall that the number of reward distribution changes is L and note that only the first ρ rounds after a change-point can belong to the set $\bar{S}(\rho)$. Thus, we have $|\bar{S}(\rho)| \leq L\rho$ and

$$\sum_{t \in \bar{S}(\rho)} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \geq \Gamma(\rho)\} \leq \sum_{t \in \bar{S}(\rho)} 1 \leq L\rho. \tag{11}$$

It remains to bound the summation over $S(\rho)$. To this end, we propose the following lemma.

Lemma 1. *Let $C \in \mathcal{C}$ be a clique. For $t \in S(\rho)$, the four events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4$ defined below cannot occur simultaneously.*

- $\mathcal{E}_1 = \{\forall a \in C - \mathcal{A}_t^*, \hat{\mu}_{t-1}(\rho, a) \le \mu_{t-1}(a) + c_{t-1}(\rho, a)\}$
- $\mathcal{E}_2 = \{\hat{\mu}_{t-1}(\rho, a_t^*) > \mu_{t-1}(a_t^*) - c_{t-1}(\rho, a_t^*)\}$
- $\mathcal{E}_3 = \{\forall a \in C - \mathcal{A}_t^*, \mu_{t-1}(a_t^*) - \mu_{t-1}(a) \ge 2c_{t-1}(\rho, a)\} \cup \{n_{t-1}(\rho, C) < \Gamma(\rho)\}$
- $\mathcal{E}_4 = \{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \ge \Gamma(\rho)\}$

Proof of Lemma 1. We prove the lemma by contradiction. Suppose the four events occur simultaneously. We have

$$\begin{aligned} \hat{\mu}_{t-1}(\rho, a_t) &\le \mu_{t-1}(a_t) + c_{t-1}(\rho, a_t) \\ \hat{\mu}_{t-1}(\rho, a_t^*) &> \mu_{t-1}(a_t^*) - c_{t-1}(\rho, a_t^*) \\ \mu_{t-1}(a_t^*) - \mu_{t-1}(a_t) &\ge 2c_{t-1}(\rho, a_t) \end{aligned}$$

which implies

$$\hat{\mu}_{t-1}(\rho, a_t^*) + c_{t-1}(\rho, a_t^*) > \mu_{t-1}(a_t^*) \ge \mu_{t-1}(a_t) + 2c_{t-1}(\rho, a_t) \ge \hat{\mu}_{t-1}(\rho, a_t) + c_{t-1}(\rho, a_t). \tag{12}$$

On the other hand, the arm selection rule of Algorithm 1 indicates

$$\hat{\mu}_{t-1}(\rho, a_t^*) + c_{t-1}(\rho, a_t^*) \le \hat{\mu}_{t-1}(\rho, a_t) + c_{t-1}(\rho, a_t). \tag{13}$$

Combining (12) and (13) leads to a contradiction and thus finishes the proof. □

Lemma 1 immediately implies that for $t \in S(\rho)$,

$$\begin{aligned} \mathbb{E}[\mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \ge \Gamma(\rho)\}] &= \Pr\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \ge \Gamma(\rho)\} \\ &= \Pr(\mathcal{E}_4) \le \Pr\left(\neg(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)\right) \le \Pr(\neg\mathcal{E}_1) + \Pr(\neg\mathcal{E}_2) + \Pr(\neg\mathcal{E}_3). \end{aligned} \tag{14}$$

We first show that $\Pr(\neg\mathcal{E}_3) = 0$ for $t \in S(\rho)$ by contradiction. Suppose $\neg\mathcal{E}_3$ occurs for $t \in S(\rho)$. Then $n_{t-1}(\rho, C) \ge \Gamma(\rho)$ and there exists an arm $a \in C - \mathcal{A}_t^*$ such that

$$\begin{aligned} \mu_{t-1}(a_t^*) - \mu_{t-1}(a) < 2c_{t-1}(\rho, a) &= \sqrt{\frac{6\log\left(|\mathcal{N}_a|^{1/3}(t-1 \wedge \rho)\right)}{O_{t-1}(\rho, a)}} \le \sqrt{\frac{6\log\left(|\mathcal{N}_a|^{1/3}(t-1 \wedge \rho)\right)}{n_{t-1}(\rho, C)}} \\ &\le \sqrt{\frac{6\log\left(|\mathcal{N}_a|^{1/3}(t-1 \wedge \rho)\right)}{\Gamma(\rho)}} = \sqrt{\frac{6\log\left(|\mathcal{N}_a|^{1/3}(t-1 \wedge \rho)\right) \cdot \Delta_{\min}^2}{6\log\left(\rho \cdot d_{\max}^{1/3} + 2\right)}} \le \Delta_{\min} \end{aligned} \tag{15}$$

where the first and the second equalities follow from the definitions of $c_{t-1}(\rho, a)$ and $\Gamma(\rho)$, respectively, the second inequality holds since $a \in C$ and in each round when an arm in C is played, the reward of a is observed, and the last inequality is due to $t - 1 \wedge \rho = \min(t-1, \rho) \le \rho$ and $d_{\max} = \max_{a' \in [K]} |\mathcal{N}_{a'}| \ge |\mathcal{N}_a|$. On the other hand, since $t \in S(\rho)$ and $a \notin \mathcal{A}_t^*$, we have

$$\mu_{t-1}(a_t^*) - \mu_{t-1}(a) = \mu_t(a_t^*) - \mu_t(a) \ge \Delta_{\min}. \tag{16}$$

Combining (15) and (16) leads to a contradiction.

Then, we turn to analyze $\Pr(\neg\mathcal{E}_2)$:

$$\begin{aligned} \Pr(\neg\mathcal{E}_2) &= \Pr\{\hat{\mu}_{t-1}(\rho, a_t^*) \le \mu_{t-1}(a_t^*) - c_{t-1}(\rho, a_t^*)\} \\ &= \Pr\left\{\mu_{t-1}(a_t^*) - \frac{1}{O_{t-1}(\rho, a_t^*)} \sum_{s=t-\rho\vee 1}^{t-1} r_s(a_t^*)\mathbb{1}\{a_s \in \mathcal{N}_{a_t^*}\} \ge \sqrt{\frac{3\log\left(|\mathcal{N}_{a_t^*}|^{1/3}(t-1 \wedge \rho)\right)}{2O_{t-1}(\rho, a_t^*)}}\right\}. \end{aligned} \tag{17}$$

Since $t \in S(\rho)$, the reward distribution of a_t^* remains fixed during rounds $[t - \rho \vee 1, t]$. Applying Hoeffding's inequality and taking the union bound over $O_{t-1}(\rho, a_t^*) \in [1, t - 1 \wedge \rho]$ gives

$$\Pr(\neg\mathcal{E}_2) \le (t - 1 \wedge \rho) \cdot \frac{1}{|\mathcal{N}_{a_t^*}|(t-1 \wedge \rho)^3} = \frac{1}{|\mathcal{N}_{a_t^*}|(t-1 \wedge \rho)^2} \le \frac{1}{(t-1 \wedge \rho)^2}.$$

Similarly, we can bound $\Pr(\neg\mathcal{E}_1)$ as

$$\begin{aligned} \Pr(\neg\mathcal{E}_1) &= \Pr\{\exists a \in C - \mathcal{A}_t^*, \hat{\mu}_{t-1}(\rho, a) > \mu_{t-1}(a) + c_{t-1}(\rho, a)\} \\ &\le \sum_{a \in C} \Pr\{\hat{\mu}_{t-1}(\rho, a) - \mu_{t-1}(a) > c_{t-1}(\rho, a)\} \\ &\le \sum_{a \in C} \frac{1}{|\mathcal{N}_a|(t-1 \wedge \rho)^2} \le \sum_{a \in C} \frac{1}{|C|(t-1 \wedge \rho)^2} = \frac{1}{(t-1 \wedge \rho)^2} \end{aligned} \tag{18}$$

where the second inequality follows from Hoeffding's inequality and the union bound, and the last inequality holds since for an arm $a \in C$, all other arms in C are neighbors of a, i.e., $C \subseteq \mathcal{N}_a$.

Finally, we combine all together. Substituting (17) and (18) into (14) gives

$$\mathbb{E}[\mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \geq \Gamma(\rho)\}] \leq \frac{2}{(t-1 \wedge \rho)^2}$$

which implies

$$\mathbb{E}\left[\sum_{t \in S(\rho)} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*, n_{t-1}(\rho, C) \geq \Gamma(\rho)\}\right] \leq \sum_{t=2}^{T} \frac{2}{(t-1 \wedge \rho)^2} \leq \sum_{t=1}^{T} \frac{2}{t^2} + \sum_{t=1}^{T} \frac{2}{\rho^2} \leq \frac{\pi^2}{3} + \frac{2T}{\rho^2}.$$

Combining this inequality with (10) and (11), we have

$$\mathbb{E}[B_2] \leq L\rho + \frac{\pi^2}{3} + \frac{2T}{\rho^2}.$$

which, together with (7) and (9), implies

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=2}^{T} \mathbb{1}\{a_t \in C - \mathcal{A}_t^*\}\right] &\leq \left(\frac{T}{\rho} + 1\right) \cdot \left(\frac{6 \log\left(\rho \cdot d_{\max}^{1/3} + 2\right)}{\Delta_{\min}^2} + 1\right) + L\rho + \frac{\pi^2}{3} + \frac{2T}{\rho^2} \\
&\leq \frac{6 \log\left(\rho \cdot d_{\max}^{1/3} + 2\right)}{\Delta_{\min}^2} \cdot \left(\frac{T}{\rho} + L\rho + 1\right) + \frac{T}{\rho} + \frac{2T}{\rho^2} + \frac{\pi^2}{3} + 1 \\
&\leq \frac{9 \log\left(\rho \cdot d_{\max}^{1/3} + 2\right)}{\Delta_{\min}^2} \cdot \left(\frac{T}{\rho} + L\rho + 1\right).
\end{aligned}$$

Substituting this inequality into (6) and using the equality $|\mathcal{C}| = \theta$, we get

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

Picking $\rho = \left\lceil \sqrt{T/L} \right\rceil$ leads to

$$\mathbb{E}[\mathrm{DR}(T)] \leq \frac{18\theta \log\left(\rho \cdot d_{\max}^{1/3} + 2\right)}{\Delta_{\min}^2} \cdot \left(\sqrt{LT} + L + 1\right) + 1 = \tilde{O}\left(\theta\sqrt{LT}\right).$$

C Proof of a Variant of Theorem 2 for Directed Feedback Graphs

As mentioned in Remark 1, we also establish a dynamic regret bound of SEASIDE for directed feedback graphs as follows.

Theorem 4. *The expected dynamic regret of SEASIDE for directed feedback graphs is upper bounded as*

$$\mathbb{E}[\mathrm{DR}(T)] \leq \left(\frac{20}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} + 2\right)\left(\frac{L}{8p} + 8\alpha p T\right) \log\left(KT^2\right) + 3.$$

Furthermore, when the number of reward distribution changes L is known in advance,[9] *by setting p optimally as $p = \sqrt{L/(64\alpha T)}$, SEASIDE achieves an $\tilde{O}(\sqrt{\alpha LT})$ dynamic regret bound for directed feedback graphs.*

Proof of Theorem 4 We start with two inequalities that hold with high probability and will be used in the analysis. First, by the following lemma

Lemma 2 (Cohen, Hazan, and Koren 2016)**.** *For directed feedback graphs, with probability $1 - 1/T^2$, AlphaSample terminates after at most $4\alpha \log\left(KT^2\right)$ rounds.*

and the union bound, with probability $1 - 1/T$, the length of each epoch in Algorithm 2 is not more than $4\alpha \log\left(KT^2\right)$, i.e.,

$$e_{\tau+1} - e_\tau \leq 4\alpha \log\left(KT^2\right), \tau = 1, 2, \ldots, m \tag{19}$$

where m denotes the number of epochs, and e_{m+1} is defined to be $T + 1$.

[9] Otherwise, we can set $p = \sqrt{1/(64\alpha T)}$ and obtain a dynamic regret bound of $\tilde{O}(L\sqrt{\alpha T})$, which is still sublinear in T.

Define $\xi_{L+1} = T + 1$ and let $1 = \xi_1 < \xi_2 < \ldots < \xi_L < T + 1$ be the change-points such that for $\ell = 1, \ldots, L$, there exists an arm whose reward distributions are different between ξ_ℓ and $\xi_\ell - 1$:

$$\exists a \in [K],\ \mathcal{D}_{\xi_\ell}(a) \neq \mathcal{D}_{\xi_\ell - 1}(a)$$

and reward distributions of all arms remain fixed during $[\xi_\ell, \xi_{\ell+1} - 1]$:

$$\forall a \in [K], \forall s, t \in [\xi_\ell, \xi_{\ell+1} - 1],\ \mathcal{D}_s(a) = \mathcal{D}_t(a).$$

Then, we decompose the T rounds as

$$[1, T] = [\xi_1, \xi_2) \cup [\xi_2, \xi_3) \cup \cdots \cup [\xi_L, \xi_{L+1}). \tag{20}$$

Let $f < g$ be any two rounds such that

- there exists $l \in [L]$ with $[f, g-1] \subseteq [\xi_l, \xi_{l+1} - 1]$;
- Algorithm 2 is reset in round f or $f = 1$;
- Algorithm 2 is reset in round g or $g = \xi_{\ell+1}$;
- Algorithm 2 is not reset during (f, g).

Let $\tau_1, \tau_2 = \tau_1 + 1, \ldots, \tau_n = \tau_1 + n - 1$ be all epochs whose beginnings are in $[f, g-1]$, i.e., $f \leq e_{\tau_i} \leq g - 1, i \in [n]$. It can be shown that $e_{\tau_1} = f$ and for $i \in [n-1]$, when computing $\hat{\mu}_{\tau_i}$ in Step 17 of Algorithm 2, $\tilde{\tau} = \tau_1 - 1$. Note that during $[e_{\tau_1}, e_{\tau_n} - 1] \subseteq [f, g-1] \subseteq [\xi_l, \xi_{l+1} - 1]$, reward distributions of all arms do not change. Applying Hoeffding's inequality and taking the union bound gives

$$\Pr\left\{\exists f, \exists i \in [n-1], \exists a \in \mathcal{A}_{\tau_i}, |\hat{\mu}_{\tau_i}(a) - \mu_f(a)| \geq \sqrt{\frac{\log(KTi)}{i}}\right\} \leq KT \sum_{i=1}^{+\infty} \frac{2}{(KTi)^2} \leq \frac{\pi^2}{3KT} \leq \frac{2}{T}$$

where we have used the equality $\tau_i - \tilde{\tau} = \tau_i - \tau_1 + 1 = i$. The above inequality implies that with probability $1 - 2/T$,

$$\forall f, \forall i \in [n-1], \forall a \in \mathcal{A}_{\tau_i}, \hat{\mu}_{\tau_i}(a) - \sqrt{\frac{\log(KTi)}{i}} \leq \mu_f(a) \leq \hat{\mu}_{\tau_i}(a) + \sqrt{\frac{\log(KTi)}{i}}. \tag{21}$$

Since in (19) and (21), the error probability $1/T$ and $2/T$ cause only $T/T = 1$ and $2T/T = 2$ expected regrets over T rounds respectively, we assume in all the following that (19) and (21) hold, and add 3 to the expected regret bound in the end of the proof.

Fix f, g. We now analyze the regret over $[f, g-1]$, which can be split into two parts:

$$\begin{aligned}
\sum_{t=f}^{g-1} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) &= \sum_{i=1}^{n-1} \sum_{t=e_{\tau_i}}^{e_{\tau_{i+1}}-1} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) + \sum_{t=e_{\tau_n}}^{g-1} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) \\
&\leq \sum_{i=1}^{n-1} \sum_{t=e_{\tau_i}}^{e_{\tau_{i+1}}-1} \mathbb{1}\{a_t \notin \mathcal{A}_f^*\} + \sum_{t=e_{\tau_n}}^{g-1} \mathbb{1}\{a_t \notin \mathcal{A}_f^*\} \\
&\leq (g - e_{\tau_n}) + \sum_{i=1}^{n-1} \sum_{t=e_{\tau_i}}^{e_{\tau_{i+1}}-1} \mathbb{1}\{a_t \notin \mathcal{A}_f^*\} \\
&\leq 4\alpha \log(KT^2) + \sum_{i=1}^{n-1} \sum_{t=e_{\tau_i}}^{e_{\tau_{i+1}}-1} \mathbb{1}\{a_t \notin \mathcal{A}_f^*\}
\end{aligned} \tag{22}$$

where recall that $\mathcal{A}_f^*$ denotes the set of all optimal arms in round f, and the last inequality follows from (19) and the fact that τ_n is the last epoch that begins in $[f, g-1]$. To bound the last summation, we propose the following lemma.

Lemma 3. *Suppose (21) hold. We have*

1. $\forall i \in [n-1]$, $\mathcal{A}_{\tau_i}$ *contains at least one optimal arm, i.e.,* $\mathcal{A}_{\tau_i} \cap \mathcal{A}_f^* \neq \varnothing$.
2. *If* $n-1 \geq \left\lceil \frac{40}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} \right\rceil + 1$, *then for* $n-1 \geq i \geq \left\lceil \frac{40}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} \right\rceil + 1$, $\mathcal{A}_{\tau_i}$ *contains only optimal arms, i.e.,* $\mathcal{A}_{\tau_i} \subseteq \mathcal{A}_f^*$.

where recall that $\Delta_{\min}$ *denotes the minimum gap of expected reward between the optimal arm and sub-optimal arms.*

Proof of Lemma 3. We first prove by induction that $\forall i \in [n-1]$,

$$\mathcal{A}_{\tau_i} \cap \mathcal{A}_f^* \neq \varnothing. \tag{23}$$

For $i = 1$, the above equality trivially holds since by Steps 1 and 10 in Algorithm 2, $\mathcal{A}_{\tau_1} = [K]$. Suppose for some $i \in [n-2]$, the above equality holds. Let $a_{\tau_i}^* \in \mathcal{A}_{\tau_i} \cap \mathcal{A}_f^*$ be an optimal arm and $\tilde{a}_{\tau_i}^* \in \arg\max_{a \in \mathcal{A}_{\tau_i}} \hat{\mu}_{\tau_i}(a)$ be an arm with the highest empirical mean reward. According to Step 18 in Algorithm 2, $\tilde{a}_{\tau_i}^* \in \mathcal{A}_{\tau_{i+1}}$. If $\tilde{a}_{\tau_i}^* \in \mathcal{A}_f^*$, then $\tilde{a}_{\tau_i}^* \in \mathcal{A}_{\tau_{i+1}} \cap \mathcal{A}_f^*$. Otherwise, $\mu_f(a_{\tau_i}^*) - \mu_f(\tilde{a}_{\tau_i}^*) > 0$. By (21), we have

$$\hat{\mu}_{\tau_i}(a_{\tau_i}^*) - \hat{\mu}_{\tau_i}(\tilde{a}_{\tau_i}^*) \geq \mu_f(a_{\tau_i}^*) - \mu_f(\tilde{a}_{\tau_i}^*) - 2\sqrt{\frac{\log(KTi)}{i}} > -2\sqrt{\frac{\log(KTi)}{i}}$$

which, together with Step 18 in Algorithm 2, implies $a_{\tau_i}^* \in \mathcal{A}_{\tau_{i+1}} \cap \mathcal{A}_f^*$. Thus, (23) holds for any $i \in [n-1]$.

We now turn to prove the second statement in Lemma 3. For $i \in [n-1]$ with $\mathcal{A}_{\tau_i} \not\subseteq \mathcal{A}_f^*$, let $a \in \mathcal{A}_{\tau_i} - \mathcal{A}_f^*$ be a sub-optimal arm and Δ_a be the gap of expected reward between arm a and the optimal arm $a_{\tau_i}^*$, i.e., $\Delta_a = \mu_f(a_{\tau_i}^*) - \mu_f(a)$. According to Step 18 in Algorithm 2, arm a is removed from $\mathcal{A}_{\tau_i}$ if (but not only if)

$$\hat{\mu}_{\tau_i}(a) \leq \hat{\mu}_{\tau_i}(a_{\tau_i}^*) - 2\sqrt{\frac{\log(KTi)}{i}}.$$

By (21), the above inequality holds if (but not only if)

$$\Delta_a = \mu_f(a_{\tau_i}^*) - \mu_f(a) \geq 4\sqrt{\frac{\log(KTi)}{i}}.$$

As $\Delta_{\min} \leq \Delta_a$, we conclude that for $i \in [n-1]$ satisfying

$$4\sqrt{\frac{\log(KTi)}{i}} \leq \Delta_{\min} \tag{24}$$

all sub-optimal arms $a \in [K] - \mathcal{A}_f^*$ are not in $\mathcal{A}_{\tau_{i+1}}$. Following the analysis in Allesiardo, Féraud, and Maillard (2017), it can be shown that when

$$i \geq \left\lceil \frac{40}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} \right\rceil$$

the inequality (24) holds. Thus, for $n-1 \geq i \geq \left\lceil \frac{40}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} \right\rceil + 1$, $\mathcal{A}_{\tau_i}$ contains no sub-optimal arms, or equivalently, only optimal arms. □

Lemma 3, together with (19) and the fact that $\forall t \in [e_{\tau_i}, e_{\tau_{i+1}-1}], a_t \in \mathcal{A}_{\tau_i}$, leads to

$$\begin{aligned}
\sum_{i=1}^{n-1} \sum_{t=e_{\tau_i}}^{e_{\tau_{i+1}}-1} \mathbb{1}\{a_t \notin \mathcal{A}_f^*\} &= \sum_{i=1}^{n-1} \sum_{t=e_{\tau_i}}^{e_{\tau_{i+1}}-1} \mathbb{1}\{a_t \in \mathcal{A}_{\tau_i}, a_t \notin \mathcal{A}_f^*\} \\
&\leq \sum_{i=1}^{n-1} \sum_{t=e_{\tau_i}}^{e_{\tau_{i+1}}-1} \mathbb{1}\{\mathcal{A}_{\tau_i} \not\subseteq \mathcal{A}_f^*\} \leq 4\alpha \log(KT^2) \sum_{i=1}^{n-1} \mathbb{1}\{\mathcal{A}_{\tau_i} \not\subseteq \mathcal{A}_f^*\} \\
&\leq 4\alpha \left\lceil \frac{40}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} \right\rceil \log(KT^2) \leq 4\alpha \left(\frac{40}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} + 1 \right) \log(KT^2).
\end{aligned}$$

Substituting the above inequality into (22), we get

$$\sum_{t=f}^{g-1} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) \leq 4\alpha \left(\frac{40}{\Delta_{\min}^2} \log \frac{7KT}{\Delta_{\min}} + 2 \right) \log(KT^2). \tag{25}$$

To proceed, we decompose the regret by using (20):

$$\mathrm{DR}(T) = \sum_{t=1}^{T} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) = \sum_{\ell=1}^{L} \sum_{t=\xi_\ell}^{\xi_{\ell+1}-1} \big(\mu_t(a_t^*) - \mu_t(a_t)\big). \tag{26}$$

Fix $\ell \in L$. Let $R_t, t \in [T]$ denote the event that in round t, Steps 10-11 of Algorithm 2 are executed and λ_ℓ denote the first round $t \in [\xi_l, \xi_{l+1} - 1]$ when event R_t occurs

$$\lambda_\ell = \min\{t \in [\xi_l, \xi_{l+1} - 1] \colon \mathbb{1}\{R_t\} = 1\}.$$

If no such t exists, we define $\lambda_\ell = \xi_{\ell+1}$. Then, we have

$$\begin{aligned}\sum_{t=\xi_\ell}^{\xi_{\ell+1}-1}\big(\mu_t(a_t^*)-\mu_t(a_t)\big) &= \sum_{t=\xi_\ell}^{\lambda_\ell-1}\big(\mu_t(a_t^*)-\mu_t(a_t)\big) + \sum_{t=\lambda_\ell}^{\xi_{\ell+1}-1}\big(\mu_t(a_t^*)-\mu_t(a_t)\big)\\ &\le \lambda_l - \xi_l + \sum_{t=\lambda_\ell}^{\xi_{\ell+1}-1}\big(\mu_t(a_t^*)-\mu_t(a_t)\big)\\ &\le \lambda_l - \xi_l + \sum_{t=\lambda_\ell}^{\xi_{\ell+1}-1}\mathbb{1}\{R_t\}\cdot 4\alpha\left(\frac{40}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+2\right)\log(KT^2)\end{aligned}$$

where the last inequality is due to (25). Taking expectation on both sides gives

$$\mathbb{E}\left[\sum_{t=\xi_\ell}^{\xi_{\ell+1}-1}\big(\mu_t(a_t^*)-\mu_t(a_t)\big)\right] \le \frac{1}{p} + 4\alpha p(\xi_{l+1}-\xi_l)\left(\frac{40}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+2\right)\log(KT^2)$$

where we have used the fact that $\mathbb{1}\{R_t\}, t\in[T]$ is a sequence of i.i.d. Bernoulli random variables with parameter p. Substituting the above inequality into (26) and noting that $\xi_{L+1}-\xi_1 = T$, we have

$$\mathbb{E}\left[\mathrm{DR}(T)\right] \le \frac{L}{p} + 4\alpha pT\left(\frac{40}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+2\right)\log(KT^2)+3 \le \left(\frac{20}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+1\right)\left(\frac{L}{8p}+8\alpha pT\right)\log(KT^2)+3$$

where the term 3 bounds the expected regret for the violation of (19) and (21).

Finally, by setting $p=\sqrt{L/(64\alpha T)}$, we get

$$\mathbb{E}\left[\mathrm{DR}(T)\right] \le \left(\frac{40}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+2\right)\sqrt{\alpha LT}\log(KT^2)+3 = \tilde{O}\big(\sqrt{\alpha LT}\big).$$

D Proof of Theorem 2

For undirected feedback graphs, the arms chosen by AlphaSample constitute an independence set. Thus, we can replace Lemma 2 in the proof of Theorem 4 at Appendix C with the following fact: For undirected feedback graphs, AlphaSample terminates after at most α rounds. Then, following the proof of Theorem 4, we can bound the dynamic regret of SEASIDE as

$$\mathbb{E}\left[\mathrm{DR}(T)\right] \le \frac{L}{p} + \alpha pT\left(\frac{40}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+2\right)+2 \le \left(\frac{10}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+\frac{1}{2}\right)\left(\frac{L}{4p}+4\alpha pT\right)+2.$$

Finally, by setting $p=\sqrt{L/(16\alpha T)}$, we get

$$\mathbb{E}\left[\mathrm{DR}(T)\right] \le \left(\frac{20}{\Delta_{\min}^2}\log\frac{7KT}{\Delta_{\min}}+1\right)\sqrt{\alpha LT}+2 = \tilde{O}\big(\sqrt{\alpha LT}\big).$$

E Proof of Theorem 3

Our proof is an adaptation of that of Auer, Gajane, and Ortner (2019) to the graph feedback setting. The main novelty of our proof is listed below.

- We propose Lemma 4 and Lemma 5, which are new and illustrate the advantages of the arm selection strategy in Algorithm 3. By utilizing these two lemmas, instead of bounding the regret for each arm separately, we can bound the regret for each clique so as to obtain a regret bound that depends on the number of cliques rather than the number of arms.
- In the analysis of the cost of playing bad arms, we decompose each epoch into time intervals where the set $\mathcal{W}_t$ does not change, and bound the cost for each interval. This decomposition is novel and crucial to deriving a bound that scales with $\sqrt{|\mathcal{W}_t|} \le \sqrt{\theta}$ rather than $\sqrt{K}$.
- In the analysis of the case that a bad arm in $\mathcal{W}_t$ becomes optimal due to the improvement of its mean reward, we utilize the properties of our proposed Algorithm 4 to bound the expected number of rounds before such improvement is detected.

We begin the proof with some lemmas that will be used. Fix a minimum clique covering $\mathcal{C} = \{C_1, C_2, \ldots, C_\theta\}$ of the feedback graph G throughout the proof. Our first two lemmas are consequences of the arm selection strategy in Algorithm 3.

Lemma 4. *For any epoch $\tau\in[m]$, any time interval $[s,t]\subseteq[e_\tau, e_{\tau+1}-1]$, and any arm a that is eligible during $[s,t]$, i.e., $\forall i\in[s,t], a\in\mathcal{E}_i$, we have*

$$O_{[s,t]}(a) \ge \lfloor (t-s+1)/\theta \rfloor. \quad (27)$$

Lemma 5. *For any epoch $\tau \in [m]$, any time interval $[s,t] \subseteq [e_\tau, e_{\tau+1}-1]$, and any two arms u, v that are eligible during $[s,t]$, i.e., $\forall i \in [s,t], u, v \in \mathcal{E}_i$, we have*

$$O_{[s,t]}(u) \geq n_{[s,t]}(C(v)) - 1 \tag{28}$$

where $C(v) \in \mathcal{C}$ denotes the clique that contains v and $n_{[s,t]}(C)$ denotes the number of times that the played arm is in clique C during $[s,t]$

$$n_{[s,t]}(C) = \sum_{i=s}^{t} \mathbb{1}\{a_i \in C\}.$$

Proof of Lemma 4. If $(t-s+1)/\theta < 1$ or $\theta = 1$, (27) holds trivially. Thus, below we only consider the case that $(t-s+1)/\theta \geq 1$ and $\theta \geq 2$. Let $N = \lfloor (t-s+1)/\theta \rfloor$ and $i_j = s + j\theta, j = 0, \ldots, N$. We show that for any $j \in [N]$ during $[i_{j-1}, i_j - 1]$, reward of arm a is observed at least once

$$O_{[i_{j-1}, i_j - 1]}(a) \geq 1. \tag{29}$$

We prove this by contradiction. Suppose there exists $j \in [N]$ such that

$$O_{[i_{j-1}, i_j - 1]}(a) = 0.$$

Let us say that a clique C is played in round t if and only if $a_t \in C$. Since the minimum clique covering $\mathcal{C}$ is comprised of θ cliques and during $[i_{j-1}, i_j - 1]$ with length $i_j - i_{j-1} = \theta$ the clique $C(a)$ is never played, there must exist a clique $C' \neq C$ that are played more than once during $[i_{j-1}, i_j - 1]$. Let $t_1, t_2 \in [i_{j-1}, i_j - 1]$ with $t_1 < t_2$ be any two rounds when C' is played. Then, $a_{t_1} \in C'$ and $a_{t_2} \in C'$. By the arm selection strategy in Algorithm 3, we have

$$\zeta_{t_2}(a_{t_2}) \leq \zeta_{t_2}(a) < i_{j-1}. \tag{30}$$

On the other hand, $a_{t_1} \in C'$ implies

$$\zeta_{t_2}(a_{t_2}) \geq t_1 \geq i_{j-1}$$

which contradicts with (30). Thus, (29) holds and

$$O_{[s,t]}(a) \geq \sum_{j=1}^{N} O_{[i_{j-1}, i_j - 1]}(a) \geq N.$$

Substituting $N = \lfloor (t-s+1)/\theta \rfloor$ into the above inequality finishes the proof. □

Proof of Lemma 5. If $n_{[s,t]}(C(v)) \leq 1$ or $u \in C(v)$, (28) trivially holds. Thus, in the following we only consider the case that $n_{[s,t]}(C(v)) > 1$ and $u \notin C(v)$. Let $N = n_{[s,t]}(C(v))$ and $s \leq i_1 < i_2 < \ldots < i_N \leq t$ be the N rounds when the played arm is in $C(v)$: $\forall j \in [N], a_{i_j} \in C(v)$. We show that for any $j \in [N-1]$, during $[i_j, i_{j+1} - 1]$, reward of arm u is observed at least once

$$\forall j \in [N-1], O_{[i_j, i_{j+1}-1]}(u) \geq 1. \tag{31}$$

We prove the above inequality by contradiction. Suppose there exists $j \in [N-1]$ such that $O_{[i_j, i_{j+1}-1]}(u) = 0$. This implies that the last round before i_{j+1} when reward of arm u is observed is not in $[i_j, i_{j+1} - 1]$

$$\zeta_{i_{j+1}}(u) < i_j.$$

By the arm selection strategy in Algorithm 3, we have

$$\zeta_{i_{j+1}}(a_{i_{j+1}}) \leq \zeta_{i_{j+1}}(u) < i_j.$$

On the other hand, $a_{i_j} \in C(v)$ implies

$$\zeta_{i_{j+1}}(a_{i_{j+1}}) \geq i_j.$$

Combining the above two inequalities, we obtain a contradiction and thus prove (31), which immediately implies

$$O_{[s,t]}(u) \geq \sum_{j=1}^{N-1} O_{[i_j, i_{j+1}-1]}(u) \geq N - 1.$$

Substituting $N = n_{[s,t]}(C(v))$ into the above inequality completes the proof. □

We proceed to propose the third lemma. Let $\mathcal{L}[s,t]$ be the number of reward distribution changes in $[s,t]$

$$\mathcal{L}[s,t] = |\{s+1 \leq i \leq t \colon \exists a \in [K], \mathcal{D}_i(a) \neq \mathcal{D}_{i-1}(a)\}|.$$

If $\mathcal{L}[s,t] = 0$, then during $[s,t]$, reward distributions of all arms are stationary and thus distance between the empirical mean reward and the expected reward of each arm can be bounded by concentration inequalities and the union bound. Formally, we introduce the following lemma, the proof of which is a simple variant of that of Lemma 5 in Auer, Gajane, and Ortner (2019) and thus omitted here.

Lemma 6. *With probability $1-2/T$, for any $[s,t] \subseteq [1,T]$ with $\mathcal{L}[s,t]=0$ and any arm $a \in [K]$,*

$$|\hat{\mu}_{[s,t]}(a) - \mu_s(a)| \leq \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}}. \tag{32}$$

Since the error probability $2/T$ causes only $T \cdot 2/T = 2$ regret over T rounds, we assume in all the following that (32) holds, and add 2 to the obtained regret bound in the end of the proof.

Based on Lemma 6, the next lemma shows that in ASG, an epoch ends and a new epoch begins only if there is an arm whose reward distribution has changed.

Lemma 7. *For all epochs $\tau \in [m-1]$, we have $\mathcal{L}[e_\tau, e_{\tau+1}-1] \geq 1$.*

Proof of Lemma 7. We prove the lemma by contradiction. Suppose there is an epoch $\tau \in [m-1]$ with $\mathcal{L}[e_\tau, e_{\tau+1}-1]=0$. Let $t = e_{\tau+1}-1$. Then, $\tau \leq m-1$ implies $t = e_{\tau+1}-1 \leq e_m - 1 \leq T$. Thus, in round t, either Step 19 or 24 of Algorithm 3 is executed. Below we analyze the two cases separately.

1) Step 19 is executed. In this case, there is a good arm $a \in \mathcal{G}_t$ such that for some $s_1, s_2, s \in [e_\tau, t]$,

$$|\hat{\mu}_{[s_1,s_2]}(a) - \hat{\mu}_{[s,t]}(a)| > \sqrt{\frac{3\log(KT)}{2O_{[s_1,s_2]}(\mathcal{G}_{s_2})}} + \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(\mathcal{G}_t)}}. \tag{33}$$

On the other hand, since $\mathcal{L}[e_\tau, t]=0$ and $s_1, s_2, s \in [e_\tau, t]$, we can apply (32) and get

$$|\hat{\mu}_{[s_1,s_2]}(a) - \mu_t(a)| \leq \sqrt{\frac{3\log(KT)}{2O_{[s_1,s_2]}(a)}}, \quad |\hat{\mu}_{[s,t]}(a) - \mu_t(a)| \leq \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}}$$

where we have used the fact that $\mathcal{D}_{s_1}(a) = \mathcal{D}_s(a) = \mathcal{D}_t(a)$. The above inequalities imply

$$|\hat{\mu}_{[s_1,s_2]}(a) - \hat{\mu}_{[s,t]}(a)| \leq |\hat{\mu}_{[s_1,s_2]}(a) - \mu_t(a)| + |\mu_t(a) - \hat{\mu}_{[s,t]}(a)| \leq \sqrt{\frac{3\log(KT)}{2O_{[s_1,s_2]}(a)}} + \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}}. \tag{34}$$

Since $a \in \mathcal{G}_t$ and $s_2 \leq t$, the construction of Algorithm 3 implies $a \in \mathcal{G}_{s_2}$. By the definition in (1), we have

$$O_{[s_1,s_2]}(a) \geq O_{[s_1,s_2]}(\mathcal{G}_{s_2}), \quad O_{[s,t]}(a) \geq O_{[s,t]}(\mathcal{G}_t)$$

which, together with (33) and (34), leads to a contradiction.

2) Step 24 is executed. In this case, there is a bad arm $a \in \mathcal{B}_t$ such that for some $s \in [e_\tau, t]$,

$$|\hat{\mu}_{[s,t]}(a) - \tilde{\mu}_\tau(a)| > \frac{\tilde{\Delta}_\tau(a)}{4} + \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}}. \tag{35}$$

Since a is a bad arm in round t, it must be eliminated from the good arm set in some previous round $t' \in [e_\tau, t)$. By Step 28 of Algorithm 3, there exists $s' \in [e_\tau, t']$ such that (3) holds and

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

Substituting this inequality into (35) gives

$$|\hat{\mu}_{[s,t]}(a) - \hat{\mu}_{[s',t']}(a)| > \frac{\sqrt{2}+1}{2}\sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}} \geq \frac{\sqrt{2}+1}{2}\sqrt{\frac{3\log(KT)}{2O_{[s',t']}(a)}} + \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}} \tag{36}$$

where the last inequality holds since $a \in \mathcal{G}_{t'}$. On the other hand, by $\mathcal{D}_{s'}(a) = \mathcal{D}_s(a) = \mathcal{D}_t(a)$ and (32), we have

$$|\hat{\mu}_{[s,t]}(a) - \hat{\mu}_{[s',t']}(a)| = |\hat{\mu}_{[s,t]}(a) - \mu_s(a) + \mu_{s'}(a) - \hat{\mu}_{[s',t']}(a)| \leq \sqrt{\frac{3\log(KT)}{2O_{[s,t]}(a)}} + \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(a)}}. \tag{37}$$

Combining (36) and (37) yields a contradiction. □

Lemma 7 immediately implies that the number of epochs can not exceed the number of reward distribution changes

$$m \leq L. \tag{38}$$

While in (2) we only partition $[1,T]$ into epochs according to the execution of ASG, to analyze the regret of ASG, we further split each epoch into time intervals where the reward distributions do not change. Specifically, for epoch $\tau \in [m]$, let $z_\tau = \mathcal{L}[e_\tau, e_{\tau+1} - 1]$ be the number of distribution changes during epoch τ. We define $\beta_{\tau,0} = e_\tau, \beta_{\tau,z_\tau+1} = e_{\tau+1}$ and denote by $\beta_{\tau,1} < \ldots < \beta_{\tau,z_\tau}$ the change-points of reward distributions in $[e_\tau, e_{\tau+1} - 1]$ such that for $i = 1, \ldots, z_\tau$, $\mathcal{L}[\beta_{\tau,i} - 1, \beta_{\tau,i}] = 1$ and for $i = 0, \ldots, z_\tau$, $\mathcal{L}[\beta_{\tau,i}, \beta_{\tau,i+1} - 1] = 0$. We further denote these time intervals with no distribution change by $I_{\tau,i} = [\beta_{\tau,i}, \beta_{\tau,i+1} - 1], i = 0, \ldots, z_\tau$. Then epoch τ can be decomposed as

$$[e_\tau, e_{\tau+1} - 1] = [\beta_{\tau,0}, \beta_{\tau,1}) \cup [\beta_{\tau,1}, \beta_{\tau,2}) \cup \cdots \cup [\beta_{\tau,z_\tau}, \beta_{\tau,z_\tau+1}) = \bigcup_{i=0}^{z_\tau} I_{\tau,i}. \tag{39}$$

We also define $\beta_{m+1,0} = \beta_{m,z_m+1}$. The reason for introducing the above decomposition is that in the proof we will often analyze the regret in an interval $I_{\tau,i}$ and then sum the derived bound over $i = 0, \ldots, z_\tau$ and $\tau = 1, \ldots, m$. The number of items in this summation is no more than $2L$, since we have

$$\sum_{\tau=1}^{m} \sum_{i=0}^{z_\tau} 1 = \sum_{\tau=1}^{m} (z_\tau + 1) = \sum_{\tau=1}^{m} z_\tau + m \leq L + m \leq 2L \tag{40}$$

where the last inequality is due to (38).

With the above notations, definitions, and lemmas, we are now ready to prove Theorem 3. Following Auer, Gajane, and Ortner (2019), we split the regret into two parts

$$\mathrm{DR}(T) = \sum_{t=1}^{T} \big(\mu_t(a_t^*) - \mu_t(a_t)\big) = \sum_{t=1}^{T} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) + \sum_{t=1}^{T} \big(\mu_t(a_t^*) - \mu_t(a_t^g)\big) \tag{41}$$

with $a_t^g \in \arg\max_{a \in \mathcal{G}_t} \mu_t(a)$ denoting the best good arm in round t. Here, the first part is the regret of the played arm with respect to the best good arm and the second part is the regret of the best good arm with respect to the optimal arm. Below, we bound the two parts separately.

We start with bounding the first part. For each epoch τ, we partition the rounds $[e_\tau, e_{\tau+1} - 1]$ into 4 sets as follows

$$[e_\tau, e_{\tau+1} - 1] = \Gamma_{\tau,1} \cup \Gamma_{\tau,2} \cup \Gamma_{\tau,3} \cup \Gamma_{\tau,4}, \quad \forall i \neq j \in [4], \Gamma_{\tau,i} \cap \Gamma_{\tau,j} = \varnothing. \tag{42}$$

where $\Gamma_{\tau,i}, i \in [4]$ is defined as follows.

- $\Gamma_{\tau,1}$ consists of all rounds when the played arm is a good arm

$$\Gamma_{\tau,1} = \{t \in [e_\tau, e_{\tau+1} - 1] \colon a_t \in \mathcal{G}_t\}.$$

- $\Gamma_{\tau,2}$ is comprised of all rounds in which the played arm is a bad arm and its regret with respect to the best good arm is not large

$$\Gamma_{\tau,2} = \Big\{t \in [e_\tau, e_{\tau+1} - 1] \colon a_t \in \mathcal{B}_t, \mu_t(a_t^g) - \mu_t(a_t) \leq 4\tilde{\Delta}_\tau(a_t)\Big\}.$$

- $\Gamma_{\tau,3}$ consists of all rounds when the the played arm is a bad arm with large regret to the best good arm and its expected reward is far away from its stored empirical mean reward

$$\Gamma_{\tau,3} = \Big\{t \in [e_\tau, e_{\tau+1} - 1] \colon a_t \in \mathcal{B}_t, \mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\tau(a_t), \tilde{\mu}_\tau(a_t) - \mu_t(a_t) > \big(\mu_t(a_t^g) - \mu_t(a_t)\big)/2\Big\}.$$

- $\Gamma_{\tau,4}$ is comprised of all rounds in which the the played arm is a bad arm with large regret to the best good arm but its expected reward is relatively close to its stored empirical mean reward

$$\Gamma_{\tau,4} = \Big\{t \in [e_\tau, e_{\tau+1} - 1] \colon a_t \in \mathcal{B}_t, \mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\tau(a_t), \tilde{\mu}_\tau(a_t) - \mu_t(a_t) \leq \big(\mu_t(a_t^g) - \mu_t(a_t)\big)/2\Big\}.$$

Based on the above definitions, we have

$$\begin{aligned}\sum_{t=1}^{T} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) = \sum_{\tau=1}^{m} \sum_{t=e_\tau}^{e_{\tau+1}-1} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) &= \sum_{\tau=1}^{m} \sum_{i=1}^{4} \sum_{t \in \Gamma_{\tau,i}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) \\ &= \sum_{i=1}^{4} \underbrace{\sum_{\tau=1}^{m} \sum_{t \in \Gamma_{\tau,i}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big)}_{B_i}.\end{aligned} \tag{43}$$

In the following, we bound B_i for each $i = 1, 2, 3, 4$.

We first analyze B_1. For $\tau \in [m]$, by (39) and the definition of $\Gamma_{\tau,1}$, we have

$$\begin{aligned}
&\sum_{t\in\Gamma_{\tau,1}} \big(\mu_t(a_t^g)-\mu_t(a_t)\big)\\
&=\sum_{i=0}^{z_\tau}\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}} \big(\mu_t(a_t^g)-\mu_t(a_t)\big)\\
&=\sum_{i=0}^{z_\tau}\sum_{C\in\mathcal{C}}\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon\, a_t\in C} \big(\mu_t(a_t^g)-\mu_t(a_t)\big) \qquad (44)\\
&=\sum_{i=0}^{z_\tau}\sum_{C\in\mathcal{C}}\left(\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon\, a_t\in C, n_{[\beta_{\tau,i},t]}(C)\le 2} \big(\mu_t(a_t^g)-\mu_t(a_t)\big)+\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon\, a_t\in C, n_{[\beta_{\tau,i},t]}(C)\ge 3} \big(\mu_t(a_t^g)-\mu_t(a_t)\big)\right)\\
&\le\sum_{i=0}^{z_\tau}\sum_{C\in\mathcal{C}}\left(2+\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon\, a_t\in C, n_{[\beta_{\tau,i},t]}(C)\ge 3} \big(\mu_t(a_t^g)-\mu_t(a_t)\big)\right)
\end{aligned}$$

where the inequality is due to $|\{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon a_t\in C, n_{[\beta_{\tau,i},t]}(C)\le 2\}|\le|\{t\ge\beta_{\tau,i}\colon a_t\in C, n_{[\beta_{\tau,i},t]}(C)\le 2\}|\le 2$ and the fact that all rewards are bounded in $[0,1]$.

For $i\in\{0,\dots,z_\tau\}$, $C\in\mathcal{C}$ and $t\in\Gamma_{\tau,1}\cap I_{\tau,i}$ with $a_t\in C, n_{[\beta_{\tau,i},t]}(C)\ge 3$, we have $a_t\in\mathcal{G}_t$ and $n_{[\beta_{\tau,i},t-1]}(C)\ge 2$. Thus, in round $t-1$ for $a=a_t$ and $s=\beta_{\tau,i}$, (3) does not hold:

$$\max_{a'\in\mathcal{G}_{t-1}}\hat{\mu}_{[\beta_{\tau,i},t-1]}(a')-\hat{\mu}_{[\beta_{\tau,i},t-1]}(a_t)\le\big(\sqrt{2}+1\big)\sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t-1]}(\mathcal{G}_{t-1})}}. \qquad (45)$$

For every arm $a\in\mathcal{G}_{t-1}$, it is eligible during $[\beta_{\tau,i},t-1]$. Since $a_t\in\mathcal{G}_t\subseteq\mathcal{G}_{t-1}$ and $a_t\in C$, we can apply Lemma 5 and get $O_{[\beta_{\tau,i},t-1]}(a)\ge n_{[\beta_{\tau,i},t-1]}(C)-1\ge 1$. So $\hat{\mu}_{[\beta_{\tau,i},t-1]}(a)$ is finite for every $a\in\mathcal{G}_{t-1}$, and we have

$$\hat{\mu}_{[\beta_{\tau,i},t-1]}(a_t^g)-\hat{\mu}_{[\beta_{\tau,i},t-1]}(a_t)\le\max_{a'\in\mathcal{G}_{t-1}}\hat{\mu}_{[\beta_{\tau,i},t-1]}(a')-\hat{\mu}_{[\beta_{\tau,i},t-1]}(a_t)\le\big(\sqrt{2}+1\big)\sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t-1]}(\mathcal{G}_{t-1})}}. \qquad (46)$$

On the other hand, since $t\le\beta_{\tau,i+1}-1$ and $\mathcal{L}[\beta_{\tau,i},\beta_{\tau,i+1}-1]=0$, by (32) we get

$$\mu_t(a_t^g)-\hat{\mu}_{[\beta_{\tau,i},t-1]}(a_t^g)\le\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t-1]}(a_t^g)}},\quad \hat{\mu}_{[\beta_{\tau,i},t-1]}(a_t)-\mu_t(a_t)\le\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t-1]}(a_t)}}$$

where we have used the fact that $\forall a\in[K],\mathcal{D}_{\beta_{\tau,i}}(a)=\mathcal{D}_t(a)$. Adding the above inequalities to (46), we obtain

$$\mu_t(a_t^g)-\mu_t(a_t)\le\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t-1]}(a_t^g)}}+\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t-1]}(a_t)}}+\big(\sqrt{2}+1\big)\sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t-1]}(\mathcal{G}_{t-1})}}\le\big(\sqrt{2}+2\big)\sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t-1]}(\mathcal{G}_{t-1})}} \qquad (47)$$

where the last inequality holds since by (1), we have $\forall a\in\mathcal{G}_{t-1}, O_{[\beta_{\tau,i},t-1]}(a)\ge O_{[\beta_{\tau,i},t-1]}(\mathcal{G}_{t-1})$.

Since any arm in $\mathcal{G}_{t-1}$ is eligible during $[\beta_{\tau,i},t-1]$ and $O_{[\beta_{\tau,i},t-1]}(\mathcal{G}_{t-1})=\min_{a\in\mathcal{G}_{t-1}}O_{[\beta_{\tau,i},t-1]}(a)$, we have $O_{[\beta_{\tau,i},t-1]}(\mathcal{G}_{t-1})\ge n_{[\beta_{\tau,i},t-1]}(C)-1$, which, together with (47), implies

$$\begin{aligned}
\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon\, a_t\in C, n_{[\beta_{\tau,i},t]}(C)\ge 3}\big(\mu_t(a_t^g)-\mu_t(a_t)\big)&=\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon\, a_t\in C, n_{[\beta_{\tau,i},t-1]}(C)\ge 2}\big(\mu_t(a_t^g)-\mu_t(a_t)\big)\\
&\le\sum_{t\in\Gamma_{\tau,1}\cap I_{\tau,i}\colon\, a_t\in C, n_{[\beta_{\tau,i},t-1]}(C)\ge 2}\big(\sqrt{2}+2\big)\sqrt{\frac{6\log(KT)}{n_{[\beta_{\tau,i},t-1]}(C)-1}}\\
&\le\sum_{h=1}^{n_{[\beta_{\tau,i},\beta_{\tau,i+1}-1]}(C)}\big(\sqrt{2}+2\big)\sqrt{\frac{6\log(KT)}{h}}\\
&\le 2\big(\sqrt{2}+2\big)\sqrt{6n_{[\beta_{\tau,i},\beta_{\tau,i+1}-1]}(C)\log(KT)}
\end{aligned}$$

where the last inequality follows from

$$\forall H \in \mathbb{N}, \sum_{h=1}^{H} \sqrt{\frac{1}{h}} \leq 2\sqrt{H}. \tag{48}$$

Substituting the above inequality into (44) gives

$$\begin{aligned}\sum_{t\in\Gamma_{\tau,1}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) &\leq \sum_{i=0}^{z_\tau}\sum_{C\in\mathcal{C}} \left(2 + 2\big(\sqrt{2}+2\big)\sqrt{6n_{[\beta_{\tau,i},\beta_{\tau,i+1}-1]}(C)\log(KT)}\right) \\ &\leq \sum_{i=0}^{z_\tau} \left(2\theta + 2\big(\sqrt{2}+2\big)\sqrt{6\theta(\beta_{\tau,i+1}-\beta_{\tau,i})\log(KT)}\right)\end{aligned} \tag{49}$$

where the second inequality is due to $|\mathcal{C}| = \theta, \sum_{C\in\mathcal{C}} n_{[\beta_{\tau,i},\beta_{\tau,i+1}-1]}(C) = \beta_{\tau,i+1} - \beta_{\tau,i}$, and

$$\forall F \in \mathbb{N}, \forall x_1, \ldots, x_F \geq 0, \sum_{f=1}^{F} \sqrt{x_f} \leq \sqrt{F\sum_{f=1}^{F} x_f}. \tag{50}$$

Summing (49) over $\tau = 1, \ldots, m$ leads to

$$\begin{aligned}B_1 = \sum_{\tau=1}^{m}\sum_{t\in\Gamma_{\tau,1}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) &\leq \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \left(2\theta + 2\big(\sqrt{2}+2\big)\sqrt{6\theta(\beta_{\tau,i+1}-\beta_{\tau,i})\log(KT)}\right) \\ &\leq 4\theta L + 4\big(\sqrt{2}+2\big)\sqrt{3\theta LT\log(KT)}\end{aligned} \tag{51}$$

where the second inequality follows from (40), (50), and

$$\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}(\beta_{\tau,i+1} - \beta_{\tau,i}) = \sum_{\tau=1}^{m}(\beta_{\tau,z_\tau+1} - \beta_{\tau,0}) \overset{(39)}{=} \sum_{\tau=1}^{m}(e_{\tau+1} - e_\tau) = e_{m+1} - e_1 \overset{(2)}{=} (T+1) - 1 = T. \tag{52}$$

We now turn to bound B_2. For $\tau \in [m]$, by the definition of $\Gamma_{\tau,2}$, we have

$$\begin{aligned}\sum_{t\in\Gamma_{\tau,2}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) &= \sum_{t\in[e_\tau,e_{\tau+1}-1]} \mathbb{1}\{a_t \in \mathcal{B}_t, \mu_t(a_t^g) - \mu_t(a_t) \leq 4\tilde{\Delta}_\tau(a_t)\}\big(\mu_t(a_t^g) - \mu_t(a_t)\big) \\ &= \sum_{a\in[K]}\sum_{t\in[e_\tau,e_{\tau+1}-1]} \mathbb{1}\{a_t = a, a \in \mathcal{B}_t, \mu_t(a_t^g) - \mu_t(a) \leq 4\tilde{\Delta}_\tau(a)\}\big(\mu_t(a_t^g) - \mu_t(a)\big) \\ &\leq \sum_{a\in[K]}\sum_{t\in[e_\tau,e_{\tau+1}-1]:\, a\in\mathcal{B}_t} \mathbb{1}\{a_t = a, \mu_t(a_t^g) - \mu_t(a) \leq 4\tilde{\Delta}_\tau(a)\} \cdot 4\tilde{\Delta}_\tau(a) \\ &\leq \sum_{a\in[K]}\sum_{t\in[e_\tau,e_{\tau+1}-1]:\, a\in\mathcal{B}_t} \mathbb{1}\{a_t = a\} \cdot 4\tilde{\Delta}_\tau(a).\end{aligned} \tag{53}$$

Let ν be the number of times that the set $\mathcal{W}_t$ changes during epoch τ. We define $\eta_0 = e_\tau, \eta_{\nu+1} = e_{\tau+1}$ and denote by $e_\tau < \eta_1 < \cdots < \eta_\nu < e_{\tau+1}$ the change points such that

$$\begin{gathered}\forall i \in [\nu], \mathcal{W}_{\eta_i} \neq \mathcal{W}_{\eta_i-1}; \\ \forall i \in \{0, \ldots, \nu\}, \forall s, t \in [\eta_i, \eta_{i+1}-1], \mathcal{W}_s = \mathcal{W}_t.\end{gathered} \tag{54}$$

For $a \in [K]$ and $t \in [e_\tau, e_{\tau+1}-1]$ with $a \in \mathcal{B}_t$, by the arm selection strategy and Step 5 in Algorithm 3, a is selected in round t, i.e., $a_t = a$, only if there is a sampling obligation $(\epsilon, n_\epsilon, s) \in \mathcal{S}_t(a)$ with $\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+$, $s \leq t$ and $O_{[s,t]}(a) \leq n_\epsilon = \lceil 1.5\epsilon^{-2}\log(KT)\rceil$. Furthermore, by Step 4 of Algorithm 3, $(\epsilon, n_\epsilon, s)$ is added into $\mathcal{S}_s(a)$ only if $a \in \mathcal{W}_s$. Thus, we have

$$\begin{aligned}\mathbb{1}\{a_t = a\} \leq& \sum_{i=0}^{\nu}\sum_{s\in[\eta_i,\eta_{i+1}-1]}\sum_{\epsilon=2^{-g}\geq\tilde{\Delta}_\tau(a)/16, g\in\mathbb{N}_+} \mathbb{1}\{a_t = a, s \leq t, O_{[s,t]}(a) \leq n_\epsilon, a \in \mathcal{W}_{\eta_i}, (\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)\} \\ \leq& \sum_{i\in[\nu]:\, a\in\mathcal{W}_{\eta_i}}\sum_{s\in[\eta_i,\eta_{i+1}-1]}\sum_{\epsilon=2^{-g}\geq\tilde{\Delta}_\tau(a)/16, g\in\mathbb{N}_+} \mathbb{1}\{a_t = a, s \leq t, O_{[s,t]}(a) \leq n_\epsilon, (\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)\}\end{aligned}$$

where $(\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)$ denotes that $(\epsilon, n_\epsilon, s)$ is added into $\mathcal{S}_s(a)$ in round s, and the second inequality is due to $\mathcal{W}_{\eta_0} = \mathcal{W}_{e_\tau} = \varnothing$. Summing the above inequality over $t \in [e_\tau, e_{\tau+1}-1]\colon a \in \mathcal{B}_t$ gives

$$
\begin{aligned}
&\sum_{t \in [e_\tau, e_{\tau+1}-1]\colon a \in \mathcal{B}_t} \mathbb{1}\{a_t = a\} \\
\leq & \sum_{t \in [e_\tau, e_{\tau+1}-1]\colon a \in \mathcal{B}_t} \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sum_{\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+} \mathbb{1}\{a_t = a, s \leq t, O_{[s,t]}(a) \leq n_\epsilon, (\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)\} \\
\leq & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sum_{\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+} \sum_{t \in [e_\tau, e_{\tau+1}-1]\colon t \geq s, a \in \mathcal{B}_t} \mathbb{1}\{a_t = a, O_{[s,t]}(a) \leq n_\epsilon, (\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)\} \\
= & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sum_{\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+} \mathbb{1}\{(\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)\} \sum_{t \in [e_\tau, e_{\tau+1}-1]\colon t \geq s, a \in \mathcal{B}_t} \mathbb{1}\{a_t = a, O_{[s,t]}(a) \leq n_\epsilon\} \\
\leq & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sum_{\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+} \mathbb{1}\{(\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)\} n_\epsilon.
\end{aligned}
$$

Taking expectation with respect to the randomness in Step 6 of Algorithm 3, we get

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t \in [e_\tau, e_{\tau+1}-1]\colon a \in \mathcal{B}_t} \mathbb{1}\{a_t = a\}\right] \\
\leq & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sum_{\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+} \mathbb{E}\left[\mathbb{1}\{(\epsilon, n_\epsilon, s) \to \mathcal{S}_s(a)\}\right] n_\epsilon \\
= & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sum_{\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+} \epsilon \sqrt{\frac{\tau}{|\mathcal{W}_s| T \log(KT)}} \left\lceil \frac{3\log(KT)}{2\epsilon^2} \right\rceil \\
\leq & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sqrt{\frac{\tau}{|\mathcal{W}_s| T \log(KT)}} \sum_{\epsilon = 2^{-g} \geq \tilde{\Delta}_\tau(a)/16, g \in \mathbb{N}_+} \left(\frac{3\log(KT)}{2\epsilon} + \epsilon\right) \\
\leq & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sum_{s \in [\eta_i, \eta_{i+1}-1]} \sqrt{\frac{\tau}{|\mathcal{W}_s| T \log(KT)}} \cdot \frac{48\log(KT)}{\tilde{\Delta}_\tau(a)} \\
= & \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sqrt{\frac{\tau \log(KT)}{|\mathcal{W}_{\eta_i}| T}} \cdot \frac{48(\eta_{i+1} - \eta_i)}{\tilde{\Delta}_\tau(a)}
\end{aligned}
$$

where the last equality is due to (54). Substituting the above inequality into (53) gives

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t \in \Gamma_{\tau,2}} \left(\mu_t(a_t^g) - \mu_t(a_t)\right)\right] &\leq \sum_{a \in [K]} \sum_{i \in [\nu]\colon a \in \mathcal{W}_{\eta_i}} \sqrt{\frac{\tau \log(KT)}{|\mathcal{W}_{\eta_i}| T}} \cdot 192(\eta_{i+1} - \eta_i) \\
&= \sum_{i \in [\nu]} \sum_{a \in \mathcal{W}_{\eta_i}} \sqrt{\frac{\tau \log(KT)}{|\mathcal{W}_{\eta_i}| T}} \cdot 192(\eta_{i+1} - \eta_i) \\
&= \sum_{i \in [\nu]} \sqrt{\frac{\tau \log(KT) |\mathcal{W}_{\eta_i}|}{T}} \cdot 192(\eta_{i+1} - \eta_i) \\
&\leq \sum_{i \in [\nu]} \sqrt{\frac{\alpha L \log(KT)}{T}} \cdot 192(\eta_{i+1} - \eta_i) \\
&\leq 192(e_{\tau+1} - e_\tau) \sqrt{\frac{\alpha L \log(KT)}{T}}
\end{aligned}
\tag{55}
$$

where the second inequality holds since $\tau \leq m \leq L$ and $\forall t \in [T], |\mathcal{W}_t| \leq \alpha$. Summing (55) over $\tau \in [m]$, we get an expected bound on B_2

$$
\mathbb{E}[B_2] = \sum_{\tau=1}^{m} \mathbb{E}\left[\sum_{t \in \Gamma_{\tau,2}} \left(\mu_t(a_t^g) - \mu_t(a_t)\right)\right] \leq \sum_{\tau=1}^{m} 192(e_{\tau+1} - e_\tau) \sqrt{\frac{\alpha L \log(KT)}{T}} = 192\sqrt{\alpha L T \log(KT)}. \tag{56}
$$

We proceed to analyze B_3:

$$\begin{aligned} B_3 &= \sum_{\tau=1}^{m} \sum_{t\in\Gamma_{\tau,3}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) = \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \sum_{t\in\Gamma_{\tau,3}\cap I_{\tau,i}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) \\ &\le 2L + \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \sum_{t\in\Gamma_{\tau,3}\cap I_{\tau,i}:\, t<\beta_{\tau,i+1}-1} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) \end{aligned} \tag{57}$$

where the inequality is due to (40) and the fact that rewards are all bounded in $[0,1]$. For $\tau \in [m]$, $i \in \{0,\dots,z_\tau\}$, and $t\in\Gamma_{\tau,3}\cap I_{\tau,i}$ with $t<\beta_{\tau,i+1}-1$, by the definition of $\Gamma_{\tau,3}$, we have $a_t\in\mathcal{B}_t$ and

$$\mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\tau(a_t), \quad \tilde{\mu}_\tau(a_t) - \mu_t(a_t) > \big(\mu_t(a_t^g) - \mu_t(a_t)\big)/2. \tag{58}$$

Since $a_t\in\mathcal{B}_t$ and in round $t<\beta_{\tau,i+1}-1\le e_{\tau+1}-1$ epoch τ does not end, by Step 23 of Algorithm 3, for $a=a_t$ and $s=\beta_{\tau,i}$, (4) does not hold:

$$|\hat{\mu}_{[\beta_{\tau,i},t]}(a_t) - \tilde{\mu}_\tau(a_t)| \le \frac{\tilde{\Delta}_\tau(a_t)}{4} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t)}}. \tag{59}$$

On the other hand, since $t\in I_{\tau,i}$ and during the time interval $I_{\tau,i}$ the reward distributions remain fixed, by (32) we get

$$|\hat{\mu}_{[\beta_{\tau,i},t]}(a_t) - \mu_t(a_t)| = |\hat{\mu}_{[\beta_{\tau,i},t]}(a_t) - \mu_{\beta_{\tau,i}}(a_t)| \le \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t)}} \tag{60}$$

Combining (58), (59), and (60) gives

$$\begin{aligned} \mu_t(a_t^g) - \mu_t(a_t) &\overset{(58)}{<} 2\big(\tilde{\mu}_\tau(a_t) - \mu_t(a_t)\big) \le 2|\tilde{\mu}_\tau(a_t) - \mu_t(a_t)| \overset{(59,60)}{\le} \frac{\tilde{\Delta}_\tau(a_t)}{2} + 4\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t)}} \\ &\overset{(58)}{\le} \frac{\mu_t(a_t^g) - \mu_t(a_t)}{8} + 2\sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(a_t)}} \end{aligned}$$

which implies $\mu_t(a_t^g) - \mu_t(a_t) < \frac{16}{7}\sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(a_t)}}$. Summing this inequality over $t\in\Gamma_{\tau,3}\cap I_{\tau,i}\colon t<\beta_{\tau,i+1}-1$ leads to

$$\begin{aligned} \sum_{t\in\Gamma_{\tau,3}\cap I_{\tau,i}:\, t<\beta_{\tau,i+1}-1} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) &\le \frac{16}{7} \sum_{t\in\Gamma_{\tau,3}\cap I_{\tau,i}:\, t<\beta_{\tau,i+1}-1} \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(a_t)}} \\ &= \frac{16}{7}\sum_{C\in\mathcal{C}} \sum_{t\in\Gamma_{\tau,3}\cap I_{\tau,i}:\, t<\beta_{\tau,i+1}-1, a_t\in\mathcal{C}} \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(a_t)}} \\ &\le \frac{16}{7}\sum_{C\in\mathcal{C}} \sum_{t\in\Gamma_{\tau,3}\cap I_{\tau,i}:\, t<\beta_{\tau,i+1}-1, a_t\in\mathcal{C}} \sqrt{\frac{6\log(KT)}{n_{[\beta_{\tau,i},t]}(C)}} \\ &\le \frac{16}{7}\sum_{C\in\mathcal{C}} \sum_{h=1}^{n_{[\beta_{\tau,i},\beta_{\tau,i+1}-1]}(C)} \sqrt{\frac{6\log(KT)}{h}} \\ &\overset{(48)}{\le} \sum_{C\in\mathcal{C}} \frac{32}{7}\sqrt{6n_{[\beta_{\tau,i},\beta_{\tau,i+1}-1]}(C)\log(KT)} \overset{(50)}{\le} 5\sqrt{6\theta(\beta_{\tau,i+1}-\beta_{\tau,i})\log(KT)}. \end{aligned}$$

Substituting the above inequality into (57) leads to

$$B_3 \le 2L + \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} 5\sqrt{6\theta(\beta_{\tau,i+1}-\beta_{\tau,i})\log(KT)} \overset{(40,50,52)}{\le} 2L + 10\sqrt{3\theta LT\log(KT)}. \tag{61}$$

Finally, we bound B_4. Similarly to (57), we decompose B_4 as

$$\begin{aligned} B_4 &= \sum_{\tau=1}^{m} \sum_{t \in \Gamma_{\tau,4}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) = \sum_{\tau=1}^{m} \sum_{i=0}^{z_\tau} \sum_{t \in \Gamma_{\tau,4} \cap I_{\tau,i}} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) \\ &\leq 2L + \sum_{\tau=1}^{m} \sum_{i=0}^{z_\tau} \sum_{t \in \Gamma_{\tau,4} \cap I_{\tau,i} :\, t < \beta_{\tau,i+1}-1} \big(\mu_t(a_t^g) - \mu_t(a_t)\big) \\ &\leq 2\theta L + \sum_{\tau=1}^{m} \sum_{i=0}^{z_\tau} \sum_{t \in \Gamma_{\tau,4} \cap I_{\tau,i} :\, \beta_{\tau,i}+\theta-1 \leq t < \beta_{\tau,i+1}-1} \big(\mu_t(a_t^g) - \mu_t(a_t)\big). \end{aligned} \tag{62}$$

For $\tau \in [m]$, $i \in \{0, \ldots, z_\tau\}$, and $t \in \Gamma_{\tau,4} \cap I_{\tau,i}$ with $\beta_{\tau,i} + \theta - 1 \leq t < \beta_{\tau,i+1} - 1$, by the definition of $\Gamma_{\tau,4}$, we have $a_t \in \mathcal{B}_t$ and

$$\mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\tau(a_t), \quad \tilde{\mu}_\tau(a_t) - \mu_t(a_t) \leq \big(\mu_t(a_t^g) - \mu_t(a_t)\big)/2. \tag{63}$$

Since a_t is a bad arm in round t, it must be eliminated from the good arm set in some previous round $t' \in [e_\tau, t)$. By Step 28 of Algorithm 3, there exists $s' \in [e_\tau, t']$ such that (3) holds for $a = a_t$ and $s = s'$, and

$$\hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t) > \big(\sqrt{2}+1\big)\sqrt{\frac{6\log(KT)}{O_{[s',t']}(\mathcal{G}_{t'})}}, \quad \tilde{\mu}_\tau(a_t) = \hat{\mu}_{[s',t']}(a_t), \quad \tilde{\Delta}_\tau(a_t) = \hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t) \tag{64}$$

with $a' \in \arg\max_{\tilde{a} \in \mathcal{G}_{t'}} \hat{\mu}_{[s',t']}(\tilde{a})$. Since during $[\beta_{\tau,i}, t]$ the reward distributions do not change, by (32) we get

$$|\hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \mu_t(a_t^g)| = |\hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \mu_{\beta_{\tau,i}}(a_t^g)| \leq \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}}.$$

Combining this inequality with (63) and (64) gives

$$\begin{aligned} \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \hat{\mu}_{[s',t']}(a_t^g) &\geq \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \hat{\mu}_{[s',t']}(a') \\ &\geq \mu_t(a_t^g) - \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}} - \hat{\mu}_{[s',t']}(a') \\ &\overset{(64)}{=} \mu_t(a_t^g) - \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}} - \tilde{\Delta}_\tau(a_t) - \tilde{\mu}_\tau(a_t) \\ &= \mu_t(a_t^g) - \mu_t(a_t) - \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}} - \tilde{\Delta}_\tau(a_t) - \big(\tilde{\mu}_\tau(a_t) - \mu_t(a_t)\big) \\ &\overset{(63)}{\geq} \frac{\mu_t(a_t^g) - \mu_t(a_t)}{4} - \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}} \end{aligned} \tag{65}$$

where the first inequality holds since $a_t^g \in \mathcal{G}_t \subseteq \mathcal{G}_{t'}$.

On the other hand, since in round $t < \beta_{\tau,i+1} - 1 \leq e_{\tau+1} - 1$, epoch τ does not end, by Step 18 of Algorithm 3, for $s_1 = s'$, $s_2 = t'$, $s = \beta_{\tau,i}$, and $a = a_t^g$, (5) does not hold. Thus, we have

$$\begin{aligned} \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \hat{\mu}_{[s',t']}(a_t^g) \leq |\hat{\mu}_{[s',t']}(a_t^g) - \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g)| &\leq \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\ &\overset{(64)}{\leq} \frac{\tilde{\Delta}_\tau(a_t)}{2\big(\sqrt{2}+1\big)} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\ &\overset{(63)}{\leq} \frac{\mu_t(a_t^g) - \mu_t(a_t)}{8\big(\sqrt{2}+1\big)} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}. \end{aligned} \tag{66}$$

Since arms in $\mathcal{G}_t$ are all eligible during $[\beta_{\tau,i}, t]$, by Lemma 4 and (1), we have $O_{[\beta_{\tau,i},t]}(\mathcal{G}_t) = \min_{a \in \mathcal{G}_t} O_{[\beta_{\tau,i},t]}(a) \geq \lfloor (t - \beta_{\tau,i} + 1)/\theta \rfloor \geq 1$, which implies that both $\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}}$ and $\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}$ are finite. Thus, we can combine (65) with (66) and get

$$\mu_t(a_t^g) - \mu_t(a_t) \leq \frac{8\big(3+\sqrt{2}\big)}{7}\left(\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}\right)$$
$$\stackrel{(1)}{\leq} \frac{16\big(3+\sqrt{2}\big)}{7}\sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \leq \frac{16\big(3+\sqrt{2}\big)}{7}\sqrt{\frac{3\log(KT)}{2\lfloor(t-\beta_{\tau,i}+1)/\theta\rfloor}}. \tag{67}$$

Summing (67) over $t \in \Gamma_{\tau,4} \cap I_{\tau,i} \colon \beta_{\tau,i}+\theta-1 \leq t < \beta_{\tau,i+1}-1$ gives

$$\sum_{t\in\Gamma_{\tau,4}\cap I_{\tau,i}\colon \beta_{\tau,i}+\theta-1\leq t<\beta_{\tau,i+1}-1}\big(\mu_t(a_t^g)-\mu_t(a_t)\big) \leq \frac{16\big(3+\sqrt{2}\big)}{7}\sum_{t\in\Gamma_{\tau,4}\cap I_{\tau,i}\colon \beta_{\tau,i}+\theta-1\leq t<\beta_{\tau,i+1}-1}\sqrt{\frac{3\log(KT)}{2\lfloor\frac{t-\beta_{\tau,i}+1}{\theta}\rfloor}}$$
$$\leq \frac{16\big(3+\sqrt{2}\big)}{7}\sum_{t=\beta_{\tau,i}+\theta-1}^{\beta_{\tau,i+1}-2}\sqrt{\frac{3\log(KT)}{2\lfloor(t-\beta_{\tau,i}+1)/\theta\rfloor}}$$
$$= \frac{16\big(3+\sqrt{2}\big)}{7}\sum_{h=\theta}^{\beta_{\tau,i+1}-\beta_{\tau,i}-1}\sqrt{\frac{3\log(KT)}{2\lfloor h/\theta\rfloor}}$$
$$\leq \frac{16\big(3+\sqrt{2}\big)}{7}\sum_{j=1}^{\lceil(\beta_{\tau,i+1}-\beta_{\tau,i})/\theta\rceil-1}\sum_{h=j\theta}^{(j+1)\theta-1}\sqrt{\frac{3\log(KT)}{2\lfloor h/\theta\rfloor}} \tag{68}$$
$$= \frac{16\big(3+\sqrt{2}\big)}{7}\sum_{j=1}^{\lceil(\beta_{\tau,i+1}-\beta_{\tau,i})/\theta\rceil-1}\theta\sqrt{\frac{3\log(KT)}{2j}}$$
$$\stackrel{(48)}{\leq} \frac{16\big(3+\sqrt{2}\big)\theta}{7}\sqrt{6\big(\lceil(\beta_{\tau,i+1}-\beta_{\tau,i})/\theta\rceil-1\big)\log(KT)}$$
$$\leq \frac{16\big(3+\sqrt{2}\big)}{7}\sqrt{6\theta(\beta_{\tau,i+1}-\beta_{\tau,i})\log(KT)}.$$

Substituting the above inequality into (62) leads to

$$B_4 \leq 2\theta L + \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\frac{16\big(3+\sqrt{2}\big)}{7}\sqrt{6\theta(\beta_{\tau,i+1}-\beta_{\tau,i})\log(KT)} \stackrel{(40,50,52)}{\leq} 2\theta L + \frac{32\big(3+\sqrt{2}\big)}{7}\sqrt{3\theta LT\log(KT)}. \tag{69}$$

By substituting (51), (56), (61), and (69) into (43), we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(\mu_t(a_t^g)-\mu_t(a_t)\big)\right] \leq 4\theta L + 4\big(\sqrt{2}+2\big)\sqrt{3\theta LT\log(KT)} + 192\sqrt{\alpha LT\log(KT)}$$
$$+ 2L + 10\sqrt{3\theta LT\log(KT)} + 2\theta L + \frac{32\big(3+\sqrt{2}\big)}{7}\sqrt{3\theta LT\log(KT)} \tag{70}$$
$$\leq 8\theta L + \frac{60\big(4+\sqrt{2}\big)}{7}\sqrt{3\theta LT\log(KT)} + 192\sqrt{\theta LT\log(KT)}.$$

Recalling the decomposition in (41), it remains to bound $\sum_{t=1}^{T}\big(\mu_t(a_t^*)-\mu_t(a_t^g)\big)$. While in each round t, there may be multiple arms that are optimal and each of them can be chosen as a_t^*, in all the following a_t^* is chosen in such way that during any interval $I_{\tau,i}$, a_t^* remains fixed

$$\forall\tau\in[m], i\in\{0,\ldots,z_\tau\}, \forall s,t\in I_{\tau,i}, a_s^* = a_t^*. \tag{71}$$

Note that this choice of a_t^* is only for regret analysis and our algorithm does not rely on it. Similarly to (42), for each epoch $\tau\in[m]$, we partition the rounds $[e_\tau, e_{\tau+1}-1]$ into 3 sets:

$$[e_\tau, e_{\tau+1}-1] = \Psi_{\tau,1}\cup\Psi_{\tau,2}\cup\Psi_{\tau,3}, \quad \forall i\neq j\in[3], \Psi_{\tau,i}\cap\Psi_{\tau,j}=\varnothing. \tag{72}$$

where $\Psi_{\tau,i}, i\in[3]$ is defined as follows.

- $\Psi_{\tau,1}$ consists of all rounds when the optimal arm is a good arm

$$\Psi_{\tau,1} = \{t\in[e_\tau, e_{\tau+1}-1]\colon a_t^*\in\mathcal{G}_t\}.$$

- $\Psi_{\tau,2}$ is comprised of all rounds in which the optimal arm is a bad arm and its expected reward is close to its stored empirical mean reward

$$\Psi_{\tau,2} = \left\{ t \in [e_\tau, e_{\tau+1}-1] \colon a_t^* \in \mathcal{B}_t, \mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) \le \tilde{\Delta}_\tau(a_t^*)/2 \right\}.$$

- $\Psi_{\tau,3}$ consists of all rounds when the optimal arm is a bad arm and its expected reward is far away from its stored empirical mean reward

$$\Psi_{\tau,3} = \left\{ t \in [e_\tau, e_{\tau+1}-1] \colon a_t^* \in \mathcal{B}_t, \mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) > \tilde{\Delta}_\tau(a_t^*)/2 \right\}.$$

Similarly to (43), we propose the following decomposition

$$\sum_{t=1}^{T} \left(\mu_t(a_t^*) - \mu_t(a_t^g)\right) = \sum_{\tau=1}^{m}\sum_{i=1}^{3}\sum_{t\in\Psi_{\tau,i}} \left(\mu_t(a_t^*) - \mu_t(a_t^g)\right) = \sum_{i=1}^{3} \underbrace{\sum_{\tau=1}^{m}\sum_{t\in\Psi_{\tau,i}} \left(\mu_t(a_t^*) - \mu_t(a_t^g)\right)}_{D_i}. \tag{73}$$

It can be shown that $D_1 = 0$ since $\mu_t(a_t^*) \ge \mu_t(a_t^g)$ trivially holds and for $t \in \Psi_{\tau,1}$, we have $a_t^* \in \mathcal{G}_t$ and $\mu_t(a_t^g) = \max_{a\in\mathcal{G}_t} \mu_t(a) \ge \mu_t(a_t^*)$. In the following, we bound D_2 and D_3 separately.

We first analyze D_2. Similarly to (62), we have

$$\begin{aligned} D_2 &= \sum_{\tau=1}^{m}\sum_{t\in\Psi\tau,2} \left(\mu_t(a_t^*) - \mu_t(a_t^g)\right) = \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\sum_{t\in\Psi_{\tau,2}\cap I_{\tau,i}} \left(\mu_t(a_t^*) - \mu_t(a_t^g)\right) \\ &\le 2\theta L + \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\sum_{t\in\Psi_{\tau,2}\cap I_{\tau,i}\colon\, \beta_{\tau,i}+\theta-1\le t<\beta_{\tau,i+1}-1} \left(\mu_t(a_t^*) - \mu_t(a_t^g)\right). \end{aligned} \tag{74}$$

For $\tau \in [m]$, $i \in \{0,\ldots,z_\tau\}$, and $t \in \Psi_{\tau,2} \cap I_{\tau,i}$ with $\beta_{\tau,i} + \theta - 1 \le t < \beta_{\tau,i+1} - 1$, by the definition of $\Psi_{\tau,2}$, a_t^* is a bad arm in round t. So it must be eliminated from the good arm set in some previous round $t' \in [e_\tau, t)$. By Step 28 of Algorithm 3, there exists $s' \in [e_\tau, t']$ such that (3) holds for $a = a_t^*$ and $s = s'$, and

$$\hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t^*) > \left(\sqrt{2}+1\right)\sqrt{\frac{6\log(KT)}{O_{[s',t']}(\mathcal{G}_{t'})}}, \quad \tilde{\mu}_\tau(a_t^*) = \hat{\mu}_{[s',t']}(a_t^*), \quad \tilde{\Delta}_\tau(a_t^*) = \hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t^*) \tag{75}$$

with $a' \in \arg\max_{\tilde{a}\in\mathcal{G}_{t'}} \hat{\mu}_{[s',t']}(\tilde{a})$. Below we consider two cases: $a' \in \mathcal{B}_t$ and $a' \in \mathcal{G}_t$.

(i) $a' \in \mathcal{B}_t$. In this case, a' is eliminated from the good arm set in some previous round $t'' \in [e_\tau, t)$ and there exists $s'' \in [e_\tau, t'']$ such that

$$\hat{\mu}_{[s'',t'']}(a'') - \hat{\mu}_{[s'',t'']}(a') > \left(\sqrt{2}+1\right)\sqrt{\frac{6\log(KT)}{O_{[s'',t'']}(\mathcal{G}_{t''})}},\ \tilde{\mu}_\tau(a') = \hat{\mu}_{[s'',t'']}(a'),\ \tilde{\Delta}_\tau(a') = \hat{\mu}_{[s'',t'']}(a'') - \hat{\mu}_{[s'',t'']}(a') \tag{76}$$

with $a'' \in \arg\max_{\tilde{a}\in\mathcal{G}_{t''}} \hat{\mu}_{[s'',t'']}(\tilde{a})$. Since in round $t'' < t < e_{\tau+1} - 1$, epoch τ does not end, by (5) and noticing that $a' \in \mathcal{G}_{t'}$ and $a' \notin \mathcal{G}_{t''+1}$ imply $t' \le t''$, we have

$$\begin{aligned} |\hat{\mu}_{[s'',t'']}(a') - \hat{\mu}_{[s',t']}(a')| &\le \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{3\log(KT)}{2O_{[s'',t'']}(\mathcal{G}_{t''})}}; \\ |\hat{\mu}_{[s'',t'']}(a'') - \hat{\mu}_{[s',t']}(a'')| &\le \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{3\log(KT)}{2O_{[s'',t'']}(\mathcal{G}_{t''})}}. \end{aligned} \tag{77}$$

Combining the above inequalities with (76) gives

$$\begin{aligned} \left(\sqrt{2}+1\right)\sqrt{\frac{6\log(KT)}{O_{[s'',t'']}(\mathcal{G}_{t''})}} &< \hat{\mu}_{[s',t']}(a'') - \hat{\mu}_{[s',t']}(a') + \sqrt{\frac{6\log(KT)}{O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{6\log(KT)}{O_{[s'',t'']}(\mathcal{G}_{t''})}} \\ &\le \sqrt{\frac{6\log(KT)}{O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{6\log(KT)}{O_{[s'',t'']}(\mathcal{G}_{t''})}} \end{aligned}$$

where the second inequality follows from the definition of a' and the fact that $a'' \in \mathcal{G}_{t''} \subseteq \mathcal{G}_{t'}$. The above inequality immediately implies

$$O_{[s'',t'']}(\mathcal{G}_{t''}) \ge 2O_{[s',t']}(\mathcal{G}_{t'}). \tag{78}$$

On the other hand, since in round t'', a_t^g is not eliminated from the good arm set, we have

$$\hat{\mu}_{[s'',t'']}(a_t^g) \geq \hat{\mu}_{[s'',t'']}(a'). \tag{79}$$

Furthermore, in round $t < e_{\tau+1} - 1$, epoch τ does not end, which implies that for $s = \beta_{\tau,i}$, $s_1 = s''$, $s_2 = t''$ and $a = a_t^g$, (5) does not hold

$$|\hat{\mu}_{[s'',t'']}(a_t^g) - \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g)| \leq \sqrt{\frac{3\log(KT)}{2O_{[s'',t'']}(\mathcal{G}_{t''})}} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}. \tag{80}$$

Finally, during $[\beta_{\tau,i}, t]$, the reward distributions remain fixed. Thus, we can apply (32) and get

$$|\hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \mu_t(a_t^g)| = |\hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \mu_{\beta_{\tau,i}}(a_t^g)| \leq \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}}. \tag{81}$$

Combining the above inequalities together, we have

$$\begin{aligned}
\mu_t(a_t^g) &\overset{(81)}{\geq} \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}} \overset{(1,80)}{\geq} \hat{\mu}_{[s'',t'']}(a_t^g) - \sqrt{\frac{3\log(KT)}{2O_{[s'',t'']}(\mathcal{G}_{t''})}} - \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\
&\overset{(79)}{\geq} \hat{\mu}_{[s'',t'']}(a') - \sqrt{\frac{3\log(KT)}{2O_{[s'',t'']}(\mathcal{G}_{t''})}} - \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\
&\overset{(77)}{\geq} \hat{\mu}_{[s',t']}(a') - \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} - \sqrt{\frac{6\log(KT)}{O_{[s'',t'']}(\mathcal{G}_{t''})}} - \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\
&\overset{(78)}{\geq} \hat{\mu}_{[s',t']}(a') - \left(\sqrt{2}+1\right)\sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} - \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}.
\end{aligned}$$

(ii) $a' \in \mathcal{G}_t$. In this case, we have the following inequality since in round $t < e_{\tau+1} - 1$, epoch τ does not end.

$$|\hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[\beta_{\tau,i},t]}(a')| \leq \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}. \tag{82}$$

On the other hand, applying Lemma 6 leads to

$$|\hat{\mu}_{[\beta_{\tau,i},t]}(a') - \mu_t(a')| = |\hat{\mu}_{[\beta_{\tau,i},t]}(a') - \mu_{\beta_{\tau,i}}(a')| \leq \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a')}} \tag{83}$$

Combining the above two inequalities gives

$$\mu_t(a_t^g) \geq \mu_t(a') \overset{(83)}{\geq} \hat{\mu}_{[\beta_{\tau,i},t]}(a') - \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a')}} \overset{(1,82)}{\geq} \hat{\mu}_{[s',t']}(a') - \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} - \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}$$

where the first inequality holds since $a_t^g \in \arg\max_{a\in\mathcal{G}_t} \mu_t(a)$ and $a' \in \mathcal{G}_t$.

Thus, in both cases we have

$$\mu_t(a_t^g) \geq \hat{\mu}_{[s',t']}(a') - \left(\sqrt{2}+1\right)\sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} - \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}. \tag{84}$$

On the other hand, by the definition of $\Psi_{\tau,2}$, we have

$$\mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) \leq \tilde{\Delta}_\tau(a_t^*)/2.$$

Combining this inequality with (75) gives

$$\hat{\mu}_{[s',t']}(a') \overset{(75)}{=} \tilde{\Delta}_\tau(a_t^*) + \tilde{\mu}_\tau(a_t^*) \geq \mu_t(a_t^*) + \frac{\tilde{\Delta}_\tau(a_t^*)}{2} \overset{(75)}{\geq} \mu_t(a_t^*) + \left(\sqrt{2}+1\right)\sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}}.$$

By the above inequality and (84), we get

$$\mu_t(a_t^*) - \mu_t(a_t^g) \leq \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}.$$

Substituting this inequality into (74) and following the analysis in (67), (68), and (69), it is easy to show that

$$D_2 \le 2\theta L + 4\sqrt{3\theta LT \log(KT)}. \tag{85}$$

We proceed to bound D_3:

$$\begin{aligned} D_3 &= \sum_{\tau=1}^{m} \sum_{t\in\Psi_{\tau,3}} \big(\mu_t(a_t^*) - \mu_t(a_t^g)\big) = \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}} \big(\mu_t(a_t^*) - \mu_t(a_t^g)\big) \\ &\le 2\theta L + \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}:\ \beta_{\tau,i}+\theta-1\le t<\beta_{\tau,i+1}-1} \big(\mu_t(a_t^*) - \mu_t(a_t^g)\big). \end{aligned} \tag{86}$$

For $\tau \in [m]$, $i \in \{0,\ldots,z_\tau\}$, and $t\in\Psi_{\tau,3}\cap I_{\tau,i}$ with $\beta_{\tau,i}+\theta-1\le t<\beta_{\tau,i+1}-1$, by the definition of $\Psi_{\tau,3}$, we have

$$a_t^* \in \mathcal{B}_t, \quad \mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) > \tilde{\Delta}_\tau(a_t^*)/2. \tag{87}$$

Since a_t^* is a bad arm in round t, it must be eliminated from the good arm set in some previous round $t' \in [e_\tau, t)$. By Step 28 of Algorithm 3, there exists $s' \in [e_\tau, t']$ such that (3) holds for $a = a_t^*$ and $s = s'$, and

$$\hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t^*) > \big(\sqrt{2}+1\big)\sqrt{\frac{6\log(KT)}{O_{[s',t']}(\mathcal{G}_{t'})}}, \quad \tilde{\mu}_\tau(a_t^*) = \hat{\mu}_{[s',t']}(a_t^*), \quad \tilde{\Delta}_\tau(a_t^*) = \hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t^*) \tag{88}$$

with $a' \in \arg\max_{\tilde{a}\in\mathcal{G}_{t'}} \hat{\mu}_{[s',t']}(\tilde{a})$. On the other hand, since in round $t < \beta_{\tau,i+1}-1 \le e_{\tau+1}-1$, epoch τ does not end, and during $[\beta_{\tau,i}, t]$ the reward distributions do not change, by Step 18 of Algorithm 3 and Lemma 6, we have

$$|\hat{\mu}_{[s',t']}(a_t^g) - \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g)| \le \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_{t})}}; \tag{89}$$

$$|\hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \mu_t(a_t^g)| = |\hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) - \mu_{\beta_{\tau,i}}(a_t^g)| \le \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(a_t^g)}}. \tag{90}$$

Combining the above inequalities together gives

$$\begin{aligned} \mu_t(a_t^*) - \mu_t(a_t^g) &\overset{(1,90)}{\le} \mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) + \tilde{\mu}_\tau(a_t^*) - \hat{\mu}_{[\beta_{\tau,i},t]}(a_t^g) + \sqrt{\frac{3\log(KT)}{2O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\ &\overset{(88,89)}{\le} \mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) + \hat{\mu}_{[s',t']}(a_t^*) - \hat{\mu}_{[s',t']}(a_t^g) + \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\ &\overset{(a_t^g\in\mathcal{G}_t,\, t>t')}{\le} \mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) + \sqrt{\frac{3\log(KT)}{2O_{[s',t']}(\mathcal{G}_{t'})}} + \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\ &\overset{(88)}{\le} \mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*) + \frac{\tilde{\Delta}_\tau(a_t^*)}{2\big(\sqrt{2}+1\big)} + \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\ &\overset{(87)}{\le} \sqrt{2}\big(\mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*)\big) + \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}}. \end{aligned} \tag{91}$$

Substituting this inequality into (86) leads to

$$\begin{aligned} D_3 &\le 2\theta L + \sqrt{2}\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}:\ \beta_{\tau,i}+\theta-1\le t<\beta_{\tau,i+1}-1} \big(\mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*)\big) \\ &\quad + \sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}:\ \beta_{\tau,i}+\theta-1\le t<\beta_{\tau,i+1}-1} \sqrt{\frac{6\log(KT)}{O_{[\beta_{\tau,i},t]}(\mathcal{G}_t)}} \\ &\le 2\theta L + 4\sqrt{3\theta LT\log(KT)} + \sqrt{2}\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau} \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}} \big(\mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*)\big) \end{aligned} \tag{92}$$

where the second inequality can be easily derived by following (67), (68), and (69).

For $\tau \in [m]$ and $i \in \{0, \ldots, z_\tau\}$, by our choice of a_t^* stated in (71), we can define $a_{\tau,i}^*$ and $\mu_{\tau,i}^*$ such that $\forall t \in I_{\tau,i}, a_t^* = a_{\tau,i}^*$ and $\mu_{\tau,i}^* = \mu_{\beta_{\tau,i}}(a_{\tau,i}^*)$. For $\tau \in [m]$ and $i \in \{0, \ldots, z_\tau\}$ with $\Psi_{\tau,3} \cap I_{\tau,i} \neq \varnothing$, we can further define

$$\epsilon_{\tau,i} = \max\ \{\epsilon\colon \epsilon = 2^{-g}, g \in \mathbb{Z}; \epsilon \leq \big(\mu_{\tau,i}^* - \tilde{\mu}_\tau(a_{\tau,i}^*)\big)/4\}, \quad n_{\tau,i} = \lceil 1.5\epsilon_{\tau,i}^{-2} \log(KT) \rceil. \tag{93}$$

Note that $\epsilon_{\tau,i} \leq 1/4$ since $\mu_{\tau,i}^* - \tilde{\mu}_\tau(a_{\tau,i}^*) \leq 1$. By the above definition, for any $t \in \Psi_{\tau,3} \cap I_{\tau,i}$, we have $2\epsilon_{\tau,i} > \big(\mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*)\big)/4$, which implies

$$\begin{aligned}
&\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\big(\mu_t(a_t^*) - \tilde{\mu}_\tau(a_t^*)\big)\\
&\leq 8\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i} = 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}\leq\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}\\
&\qquad + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}\\
&\leq 8\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\sqrt{3\theta|I_{\tau,i}|\log(KT)} + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}\\
&\overset{(50)}{\leq} 8\sqrt{6\theta LT\log(KT)} + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}, |I_{\tau,i}|\leq 2\theta n_{\tau,i}} 2\theta n_{\tau,i}\epsilon_{\tau,i}\\
&\qquad + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}, |I_{\tau,i}|> 2\theta n_{\tau,i}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}\\
&\overset{(93)}{\leq} 8\sqrt{6\theta LT\log(KT)} + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}, |I_{\tau,i}|\leq 2\theta n_{\tau,i}} \Big(\frac{3\theta\log(KT)}{\epsilon_{\tau,i}} + 2\theta\epsilon_{\tau,i}\Big)\\
&\qquad + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}, |I_{\tau,i}|> 2\theta n_{\tau,i}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}\\
&\leq 8\sqrt{6\theta LT\log(KT)} + 8\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\sqrt{3\theta|I_{\tau,i}|\log(KT)} + 16\theta\sum_{\tau=1}^{m}\sum_{i=0}^{z_\tau}\frac{1}{4}\\
&\qquad + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}, |I_{\tau,i}|> 2\theta n_{\tau,i}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}\\
&\overset{(50)}{\leq} 16\sqrt{6\theta LT\log(KT)} + 8\theta L + 8\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}, |I_{\tau,i}|> 2\theta n_{\tau,i}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}.
\end{aligned} \tag{94}$$

It remains to bound the last summation in the above inequality. To this end, we introduce $R_{\tau,i}$, which is defined as the total future expected contributions of $\epsilon_{\tau',i'}$ starting from $\beta_{\tau,i}$, i.e., the expectation of summation of contributions over τ', i', and $t \in \Psi_{\tau',3} \cap I_{\tau',i'}$ such that $t \geq \beta_{\tau,i}, \Psi_{\tau',3} \cap I_{\tau',i'} \neq \varnothing, \epsilon_{\tau',i'} > \sqrt{3\theta(\log(KT))/|I_{\tau',i'}|}, |I_{\tau',i'}| > 2\theta n_{\tau',i'}$:

$$R_{\tau,i} = \mathbb{E}\left[\sum_{\tau'=1}^{m}\sum_{i'\in\{0,\ldots,z_{\tau'}\}\colon \Psi_{\tau',3}\cap I_{\tau',i'}\neq\varnothing, \epsilon_{\tau',i'}>\sqrt{3\theta(\log(KT))/|I_{\tau',i'}|}, |I_{\tau',i'}|>2\theta n_{\tau',i'}}\ \sum_{t\in\Psi_{\tau',3}\cap I_{\tau',i'}\colon t\geq\beta_{\tau,i}}\epsilon_{\tau',i'}\ \Big|\ \beta_{\tau,i}\right].$$

Since $\beta_{1,0} = 1$, we have

$$\mathbb{E}\left[\sum_{\tau=1}^{m}\sum_{i\in\{0,\ldots,z_\tau\}\colon \Psi_{\tau,3}\cap I_{\tau,i}\neq\varnothing, \epsilon_{\tau,i}>\sqrt{3\theta(\log(KT))/|I_{\tau,i}|}, |I_{\tau,i}|>2\theta n_{\tau,i}}\ \sum_{t\in\Psi_{\tau,3}\cap I_{\tau,i}}\epsilon_{\tau,i}\right] = R_{1,0}. \tag{95}$$

To proceed, we introduce some notations. Let $1 = \xi_1 < \ldots < \xi_L$ be all rounds when there is an arm whose reward distribution changes and define $\xi_{L+1} = T + 1$. Then, $[1, T]$ can be divided into intervals $J_\ell = [\xi_\ell, \xi_{\ell+1} - 1], \ell \in [L]$:

$$[1, T] = \bigcup_{\ell=1}^{L} J_\ell.$$

It can be shown that each $I_{\tau,i} = [\beta_{\tau,i}, \beta_{\tau,i+1} - 1]$ is a subset of some J_ℓ. Let $\rho(\tau, i)$ be such that $I_{\tau,i} \subseteq J_{\rho(\tau,i)} = [\xi_{\rho(\tau,i)}, \xi_{\rho(\tau,i)+1} - 1]$. By the definitions of $I_{\tau,i}$ and J_ℓ, $\rho(\tau, i)$ is deterministic conditioned on $\beta_{\tau,i}$.

Below we prove by backward induction that

$$R_{\tau,i} \leq \sum_{\ell=\tau}^{L} \sqrt{\frac{\theta T \log (KT)}{\ell}} + \sqrt{3\theta(2L - \rho(\tau, i) - \tau)(T + 1 - \beta_{\tau,i}) \log (KT)}. \tag{96}$$

Proof of (96). The inequality trivially holds after all epochs when $\beta_{\tau,i} = T + 1$.

Consider τ and i such that $\Psi_{\tau,3} \cap I_{\tau,i} \neq \varnothing, \epsilon_{\tau,i} > \sqrt{3\theta(\log (KT))/|I_{\tau,i}|}, |I_{\tau,i}| > 2\theta n_{\tau,i}$. For $t \in \Psi_{\tau,3} \cap I_{\tau,i}$, if during $[\beta_{\tau,i}, t]$, the reward of arm $a^*_{\tau,i}$ is observed not less than $n_{\tau,i}$ times, i.e., $O_{[\beta_{\tau,i},t]}(a^*_{\tau,i}) \geq n_{\tau,i}$. Then, by (32), (87), and (93), we have

$$\begin{aligned}
\hat{\mu}_{[\beta_{\tau,i},t]}(a^*_{\tau,i}) - \tilde{\mu}_\tau(a^*_{\tau,i}) &\overset{(32)}{\geq} \mu_{\beta_{\tau,i}}(a^*_{\tau,i}) - \tilde{\mu}_\tau(a^*_{\tau,i}) - \sqrt{\frac{3 \log (KT)}{2O_{[\beta_{\tau,i},t]}(a^*_{\tau,i})}} \\
&\geq \mu^*_{\tau,i} - \tilde{\mu}_\tau(a^*_{\tau,i}) + \sqrt{\frac{3 \log (KT)}{2n_{\tau,i}}} - 2\sqrt{\frac{3 \log (KT)}{2n_{\tau,i}}} \\
&\overset{(93)}{\geq} \mu^*_{\tau,i} - \tilde{\mu}_\tau(a^*_{\tau,i}) + \sqrt{\frac{3 \log (KT)}{2n_{\tau,i}}} - 2\epsilon_{\tau,i} \\
&\overset{(93)}{\geq} \frac{\mu^*_{\tau,i} - \tilde{\mu}_\tau(a^*_{\tau,i})}{2} + \sqrt{\frac{3 \log (KT)}{2n_{\tau,i}}} \\
&\overset{(87)}{>} \frac{\tilde{\Delta}_\tau(a^*_{\tau,i})}{4} + \sqrt{\frac{3 \log (KT)}{2n_{\tau,i}}} \geq \frac{\tilde{\Delta}_\tau(a^*_{\tau,i})}{4} + \sqrt{\frac{3 \log (KT)}{2O_{[\beta_{\tau,i},t]}(a^*_{\tau,i})}}
\end{aligned}$$

which implies that (4) holds for $s = \beta_{\tau,i}$, and epoch τ ends.

On the other hand, in round $s \in \Psi_{\tau,3} \cap I_{\tau,i}$, since $\frac{1}{4} \geq \epsilon_{\tau,i} \geq \frac{\mu_s(a^*_s) - \tilde{\mu}_\tau(a^*_s)}{8} \geq \frac{\tilde{\Delta}_\tau(a^*_s)}{16} = \frac{\tilde{\Delta}_\tau(a^*_{\tau,i})}{16}$, by the properties of Algorithm 4 and Steps 4–8 of Algorithm 3, with probability $\epsilon_{\tau,i}\sqrt{\tau/\left(|\mathcal{W}_s|\, T \log (KT)\right)} \geq \epsilon_{\tau,i}\sqrt{\tau/\left(\theta T \log (KT)\right)}$, the sampling obligation $(\epsilon_{\tau,i}, n_{\tau,i}, s)$ is added into $\mathcal{S}_s(a)$ for an arm $a \in \mathcal{N}_{a^*_{\tau,i}} \cap \mathcal{W}_s$ with $\tilde{\Delta}_\tau(a^*_{\tau,i}) \geq \tilde{\Delta}_\tau(a)$. Once such sampling obligation is added into $\mathcal{S}_s(a)$, after at most $\theta n_{\tau,i}$ rounds the reward of arm $a^*_{\tau,i}$ is observed not less than $n_{\tau,i}$ times, which leads to the ending of epoch τ. Thus, denoting by $R'_{\tau,i}$ the expected contributions of $\epsilon_{\tau,i}$ within $I_{\tau,i}$ when $t \in \Psi_{\tau,3}$:

$$R'_{\tau,i} = \mathbb{E}\left[\sum_{t \in \Psi_{\tau,3} \cap I_{\tau,i}} \epsilon_{\tau,i} \,\middle|\, \beta_{\tau,i}\right]$$

and defining $p_\tau = \sqrt{\tau/\left(\theta T \log (KT)\right)}$, we have

$$\begin{aligned}
R'_{\tau,i} &\leq \epsilon_{\tau,i} \left(\sum_{h=1}^{\xi_{\rho(\tau,i)+1} - \beta_{\tau,i} - \theta n_{\tau,i}} \left(1 - \epsilon_{\tau,i}\sqrt{\tau/\left(\theta T \log (KT)\right)}\right)^h + \theta n_{\tau,i} \right) \\
&= \epsilon_{\tau,i} \left(\frac{(1 - \epsilon_{\tau,i} p_\tau)\left(1 - (1 - \epsilon_{\tau,i} p_\tau)^{\xi_{\rho(\tau,i)+1} - \beta_{\tau,i} - \theta n_{\tau,i}}\right)}{\epsilon_{\tau,i} p_\tau} + \theta n_{\tau,i} \right) \\
&\leq \frac{1 - (1 - \epsilon_{\tau,i} p_\tau)^{\xi_{\rho(\tau,i)+1} - \beta_{\tau,i} - \theta n_{\tau,i}}}{p_\tau} + \theta n_{\tau,i}\epsilon_{\tau,i} \leq \frac{1 - (1 - \epsilon_{\tau,i} p_\tau)^{\xi_{\rho(\tau,i)+1} - \beta_{\tau,i} - \theta n_{\tau,i}}}{p_\tau} + \frac{3\theta \log (KT)}{\epsilon_{\tau,i}}
\end{aligned}$$

where the last inequality holds since by (93), $n_{\tau,i} \leq 1.5\epsilon^{-2}_{\tau,i} \log (KT) + 1 \leq 3\epsilon^{-2}_{\tau,i} \log (KT)$. Furthermore, denoting by $q_{\tau,i}$ the probability that epoch τ does not end within $I_{\tau,i}$, we have

$$q_{\tau,i} \leq (1 - \epsilon_{\tau,i} p_\tau)^{\xi_{\rho(\tau,i)+1} - \beta_{\tau,i} - \theta n_{\tau,i}}.$$

Note that $\rho(\tau, i+1) = \rho(\tau, i) + 1, |I_{\tau,i}| = \beta_{\tau,i+1} - \beta_{\tau,i}$ if epoch τ does not end within $I_{\tau,i}$, and $\rho(\tau+1, 0) = \rho(\tau, i), |I_{\tau,i}| = \beta_{\tau+1,0} - \beta_{\tau,i}$ otherwise.

By the above two inequalities and induction, we obtain

$$
\begin{aligned}
R_{\tau,i} &\le R'_{\tau,i} + q_{\tau,i} R_{\tau,i+1} + (1 - q_{\tau,i}) R_{\tau+1,0} \\
&\le \frac{1 - q_{\tau,i}}{p_\tau} + q_{\tau,i}\sqrt{3\theta(\beta_{\tau,i+1} - \beta_{\tau,i}) \log(KT)} + (1 - q_{\tau,i})\sqrt{3\theta(\beta_{\tau+1,0} - \beta_{\tau,i}) \log(KT)} \\
&\quad + q_{\tau,i}\left(\sum_{\ell=\tau}^{L} \sqrt{\frac{\theta T \log(KT)}{\ell}} + \sqrt{3\theta(2L - \rho(\tau, i) - 1 - \tau)(T + 1 - \beta_{\tau,i+1}) \log(KT)}\right) \\
&\quad + (1 - q_{\tau,i})\left(\sum_{\ell=\tau+1}^{L} \sqrt{\frac{\theta T \log(KT)}{\ell}} + \sqrt{3\theta(2L - \rho(\tau, i) - \tau - 1)(T + 1 - \beta_{\tau+1,0}) \log(KT)}\right) \\
&= (1 - q_{\tau,i})\sqrt{\frac{\theta T \log(KT)}{\tau}} + q_{\tau_i}\sum_{\ell=\tau}^{L}\sqrt{\frac{\theta T \log(KT)}{\ell}} + (1 - q_{\tau,i})\sum_{\ell=\tau+1}^{L}\sqrt{\frac{\theta T \log(KT)}{\ell}} \\
&\quad + q_{\tau,i}\left(\sqrt{3\theta(\beta_{\tau,i+1} - \beta_{\tau,i}) \log(KT)} + \sqrt{3\theta(2L - \rho(\tau, i) - 1 - \tau)(T + 1 - \beta_{\tau,i+1}) \log(KT)}\right) \\
&\quad + (1 - q_{\tau,i})\left(\sqrt{3\theta(\beta_{\tau+1,0} - \beta_{\tau,i}) \log(KT)} + \sqrt{3\theta(2L - \rho(\tau, i) - \tau - 1)(T + 1 - \beta_{\tau+1,0}) \log(KT)}\right) \\
&\le \sum_{\ell=\tau}^{L}\sqrt{\frac{\theta T \log(KT)}{\ell}} + q_{\tau_i}\sqrt{3\theta \log(KT)}\sqrt{(2L - \rho(\tau, i) - \tau)(T + 1 - \beta_{\tau,i})} \\
&\quad + (1 - q_{\tau,i})\sqrt{3\theta \log(KT)}\sqrt{(2L - \rho(\tau, i) - \tau)(T + 1 - \beta_{\tau,i})} \\
&= \sum_{\ell=\tau}^{L}\sqrt{\frac{\theta T \log(KT)}{\ell}} + \sqrt{3\theta(2L - \rho(\tau, i) - \tau)(T + 1 - \beta_{\tau,i}) \log(KT)}
\end{aligned}
$$

where we have used $\epsilon_{\tau,i} > \sqrt{3\theta(\log(KT))/|I_{\tau,i}|}$, and the last inequality is due to $\sqrt{x} + \sqrt{by} \le \sqrt{(b+1)(x+y)}$. □

(96) immediately implies

$$R_{1,0} \le \sum_{\ell=1}^{L}\sqrt{\big(\theta T \log(KT)\big)/\ell} + \sqrt{6\theta LT \log(KT)} \le (2 + \sqrt{6})\sqrt{\theta LT \log(KT)}.$$

Combining this inequality with (92), (94), and (95) gives

$$
\begin{aligned}
\mathbb{E}[D_3] &\le 2\theta L + 4\sqrt{3\theta LT \log(KT)} + \sqrt{2}\left(24\sqrt{6\theta LT \log(KT)} + 8\theta L + 16\sqrt{\theta LT \log(KT)}\right) \\
&= 2\theta L + 8\sqrt{2}\theta L + \big(52\sqrt{3} + 16\sqrt{2}\big)\sqrt{\theta LT \log(KT)}.
\end{aligned}
$$

We end the proof with combining all things together. By (73), (85), and the above inequality, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(\mu_t(a_t^*) - \mu_t(a_t^g)\big)\right] \le 4\theta L + 8\sqrt{2}\theta L + \big(56\sqrt{3} + 16\sqrt{2}\big)\sqrt{\theta LT \log(KT)}$$

which, together with (41) and (70), leads to

$$
\begin{aligned}
\mathbb{E}[\mathrm{DR}(T)] &\le 8\theta L + \frac{60\big(4 + \sqrt{2}\big)}{7}\sqrt{3\theta LT \log(KT)} + 192\sqrt{\theta LT \log(KT)} \\
&\quad + 4\theta L + 8\sqrt{2}\theta L + \big(56\sqrt{3} + 16\sqrt{2}\big)\sqrt{\theta LT \log(KT)} + 2 \\
&\le (12 + 8\sqrt{2})\theta L + \frac{60\sqrt{6} + 632\sqrt{3} + 112\sqrt{2} + 1344}{7}\sqrt{\theta LT \log(KT)} + 2 \\
&\le 24\theta L + 393\sqrt{\theta LT \log(KT)} + 2
\end{aligned}
$$

where the term 2 is introduced to bound the expected regret for the violation of (32) in Lemma 6. Finally, since rewards are all bounded in $[0, 1]$, we have $\mathbb{E}[\mathrm{DR}(T)] \le T$ and

$$\mathbb{E}[\mathrm{DR}(T)] \le \min\big(24\theta L + 393\sqrt{\theta LT \log(KT)} + 2, T\big) = O\big(\sqrt{\theta LT \log(KT)}\big)$$

where the equality holds since $\theta L = \sqrt{\theta L}\sqrt{\theta L} \le \sqrt{\theta LT}$ if $\theta L \le T$, and $T = \sqrt{T}\sqrt{T} \le \sqrt{\theta LT}$ otherwise.