Proceedings of Machine Learning Research vol 99:1–22, 2019 32nd Annual Conference on Learning Theory

Adaptively Tracking the Best Bandit Arm with an Unknown Number of Distribution Changes

Peter Auer AUER@UNILEOBEN.AC.AT
Pratik Gajane PRATIK.GAJANE@UNILEOBEN.AC.AT
Ronald Ortner RONALD.ORTNER@UNILEOBEN.AC.AT
Montanuniversität Leoben

Editors: Alina Beygelzimer and Daniel Hsu

[1] **Abstract**

We consider the variant of the stochastic multi-armed bandit problem where the stochastic reward distributions may change abruptly several times. In contrast to previous work, we are able to achieve (nearly) optimal mini-max regret bounds *without knowing the number of changes*. For this setting, we propose an algorithm called ADSWITCH and provide performance guarantees for the regret evaluated against the optimal non-stationary policy. Our regret bound is the first optimal bound for an algorithm that is not tuned with respect to the number of changes.

Keywords: multi-armed stochastic bandits, non-stationary rewards, switching bandits

1. Introduction

The classical multi-armed bandit (MAB) problem is the simplest setting that gives rise to the exploration-exploitation dilemma inherent to all reinforcement learning problems (see Bubeck and Cesa-Bianchi, 2012, for a survey). In this setup, a learner has access to a number of available actions, also called "arms" in reference to the arm of a slot machine or a one-armed bandit. The learner has to repeatedly select one of these arms, which yields a reward generated from the unknown reward process of the selected arm. The learner's aim is to maximize the sum of the gathered rewards. In the usual stochastic MAB problem, the reward process for an arm is assumed to be a distribution which remains stationary. In this article, though, we consider the stochastic MAB problem with non-stationary reward distributions. Following Garivier and Moulines (2011), we call this the *switching bandits* problem. As a motivation, consider the problem of real-time content optimization of websites which aims to serve targeted and relevant content to individuals. In order to do so, the website needs to learn which content (represented by an arm of the MAB) the users are most likely to be interested in. The user interest in the content of a website (for example, news) is likely to vary over time. For additional motivating examples and practical applications of this problem setting, see Garivier and Moulines (2011), Hartland et al. (2006), Koulouriotis and Xanthopoulos (2008), and the references therein.

1. Some preliminary results have been presented at EWRL 2018 (Auer et al., 2018).

1.1. Related work

Non-stationary MAB problems where reward distributions vary over time have been previously studied in the literature. Sometimes, there are additional assumptions on the process generating the changes like in the general restless bandits setting (Ortner et al., 2014) or special cases as considered by Slivkins and Upfal (2008). In some cases, the learner is allowed to collect additional side-information (Yu and Mannor, 2009). For the setting we consider here, several approaches have been proposed. These range from modifying algorithms for the standard stochastic MAB setting like UCB (Kocsis and Szepesvári, 2006) to evolutionary algorithms (Koulouriotis and Xanthopoulos, 2008). The algorithm we introduce in this paper has more in common with (Hartland et al., 2006) where a change point detection procedure is suggested.

We note that algorithms that work in the stochastic as well in the adversarial setting (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016) usually also need to detect changes and are hence related. But the regret for these algorithms is still defined in respect to the single best arm, while we are interested in the regret in respect to best arm in each time step.

Such regret bounds have already been achieved by Auer et al. (2002) for EXP3.S, a variant of EXP3. If the number of changes L is known in advance, EXP3.S can be tuned to obtain a regret bound of $\tilde{O}\left(\sqrt{KLT}\right)$ after T steps, where K is the number of arms. This is mini-max optimal in all parameters up to $\log T$ factors, which are hidden in the $\tilde{O}$-notation.[2] It is worth mentioning that EXP3.S works also in the more general adversarial setting, where rewards are not generated by distributions but are set arbitrarily. The regret is then defined in respect to the best strategy in hindsight that may change the selected arm only a fixed number of times.

More recently, Garivier and Moulines (2011) have shown regret bounds for the discounted-UCB algorithm of (Kocsis and Szepesvári, 2006) as well as for a sliding window adaptation of UCB. Regret bounds for an elimination algorithm with restarts that is similar to our approach have been proven by Allesiardo et al. (2017). Algorithms that work even for stochastic contextual bandits have been analyzed by Luo et al. (2018). As for EXP3.S, all these algorithms can be tuned to obtain bounds optimal with respect to L and T, provided that the number of changes L is known in advance.

We note that there are also regret bounds that do not depend on the number of changes but the total variation of change V (Besbes et al., 2014; Luo et al., 2018). If V is known in advance, these algorithms can be tuned to achieve $\tilde{O}\left((KV)^{1/3}T^{2/3}\right)$ regret, which is optimal.

In this paper, we propose the algorithm ADSWITCH, which —unlike the mentioned approaches— does not need to know the number of changes L in advance. Still, we can prove an optimal regret bound $\tilde{O}\left(\sqrt{KLT}\right)$ for ADSWITCH.

1.2. Outline

Section 2 gives the problem statement and our main result. The algorithm ADSWITCH is described in Section 3, followed by Section 4, which gives part of the proof of the regret bound. The detailed proofs can be found in the appendix.

2. The regret bound cannot be improved even if the gap between the mean reward of the best arm and the mean rewards of the other arms is lower bounded by a constant, say $1/4$.

2. Problem statement and result

For notational convenience we assume that all rewards are bounded in $[0, 1]$. We denote a stochastic bandit problem with changing reward distributions by the mean rewards $\mu_t(a)$ of the arms $a = 1, \dots, K$ at times $t = 1, \dots, T$. The mean rewards $\mu_t(a)$ are unknown to the algorithm. At each time t, an algorithm selects some arm a_t and receives reward r_t with mean $\mathbb{E}[r_t] = \mu_t(a_t)$. The goal of the algorithm is to achieve low expected regret R against an omniscient policy that at each time chooses the arm with maximal mean reward,

$$R = \sum_{t=1}^{T} \max_a \mu_t(a) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_t(a_t)\right].$$

We assume that the horizon T is known, an unknown horizon can be handled by the doubling trick (Besson and Kaufmann, 2018).

For meaningful results the amount of change in the means $\mu_t(a)$ needs to be taken into account. As such a measure of change we consider the total number of changes

$$L = \#\{1 \le t \le T | \exists a : \mu_{t-1}(a) \neq \mu_t(a)\}.$$

(For notational convenience we define $\mu_0(a) = 0$ for all arms a.)

As mentioned above, the best known regret bound in terms of L is $\tilde{O}\left(\sqrt{KLT}\right)$ (see for example Auer et al., 2002), which matches the lower bound up to logarithmic factors. This bound was previously achieved for known L. In this paper, we show that the same bound can be achieved without knowing L, using an algorithm that adapts to the observed change.

Theorem 1 *For a switching bandit problem with K arms, L changes, and horizon T, the expected regret of* ADSWITCH *(introduced in Section 3 below) is upper bounded by*

$$C\sqrt{KLT\log T}$$

for a suitable constant C.

2.1. The difficulty of unknown L and how to deal with it

In this section, we give some intuition, why it is significantly harder to achieve optimal regret bounds when the number of changes is unknown, and how our algorithm deals with this difficulty.

2.1.1. CALCULATING THE SAMPLING RATE FOR INFERIOR ARMS

Assume that an algorithm has found an inferior arm a that is Δ-worse than the best arm. Then the algorithm has to safeguard against a change of arm a that would make it the best arm. When the total number of changes L is known in advance, then this is easily done by sampling arm a with probability $p = \sqrt{L/(KT)}/\Delta$. When a change is detected, then the algorithm restarts.

The total cost for this sampling is $pT\Delta = \sqrt{LT/K}$ when there is no change (since the algorithm loses the amount of Δ each time the inferior arm a is sampled). Summing over all inferior arms, this contributes $\sqrt{KLT}$ to the regret.

If arm a changes by amount $\epsilon > \Delta$, then (up to logarithmic factors) $1/\epsilon^2$ samples from arm a are sufficient to detect the change. Thus, because of the sampling rate p, the change is detected after

$1/(p\epsilon^2)$ time steps, and the regret during these time steps is at most $\epsilon/(p\epsilon^2) = \Delta\sqrt{KT/L}/\epsilon < \sqrt{KT/L}$. Summing over the changes gives again a contribution to the regret of $\sqrt{KLT}$.

The sampling rate p is chosen optimally to trade off the regret caused by sampling and the regret until a change is detected, by solving $pT\Delta K = L/(p\Delta)$ for p.

2.1.2. ADAPTING THE SAMPLING RATE IN RESPECT TO THE OBSERVED CHANGES

If the number of changes L is not known, the sampling rate p cannot be set so easily. A first attempt is to count the number of changes so far, ℓ, and set $p_\ell = \sqrt{\ell/(KT)}/\Delta$ accordingly (initially $\ell = 1$). But this does not work, since in the beginning the sampling rate $p_1 = \sqrt{1/(KT)}/\Delta$ is small and quick changes will not be detected: Let $\mu = 1/2$ be the mean reward of arm a so far and let $\Delta = 1/8$ (the mean reward of the best arm is $5/8$). Now let the mean reward of arm a alternate every W steps between $\mu + \epsilon$ and $\mu - \epsilon$ for $\epsilon = 1/4$. Since arm a is sampled only every $1/p_1 \sim T^{1/2}$ steps, the received rewards appear random with mean μ, when $W \leq T^{1/4}$. Thus no change is detected and linear regret in incurred.

Increasing the initial sampling rate p_1 would allow to detect such changes, but it would also increase the sampling cost in the case when there are few changes, such that the optimal regret bound is not achieved.

The phenomenon of increased regret rates when the number of changes is under- or overestimated can already been seen in the regret bounds for EXP3.S in (Auer et al., 2002) and its variant SHIFTBAND (Auer, 2002, Theorem 2) with the regret bound

$$\tilde{O}\left(\left(\sqrt{L_0} + L/\sqrt{L_0}\right)\sqrt{KT}\right),$$

where L_0 is a tuning parameter, interpreted as an a priori estimate for the number of changes. If indeed $L = L_0$, then the bound becomes $\tilde{O}\left(\sqrt{KLT}\right)$ and is optimal. For $L_0 = 1$, the bound is $\tilde{O}\left(L\sqrt{KT}\right)$ and gives linear regret if $L \geq \sqrt{T}$. For $L = 1$, the bound gives $\tilde{O}\left(\sqrt{KL_0T}\right)$ and is sub-optimal for large L_0.

2.1.3. CONSECUTIVE SAMPLING

The failure to detect changes as described in the previous section is caused by changes between the times a sample is drawn for the inferior arm. This can be avoided by drawing consecutive samples: Let ℓ again be the number of changes observed so far. To check for a change of size $\epsilon > \Delta$, with probability $p_\epsilon = \epsilon\sqrt{\ell/(KT)}$ the algorithm draws $n_\epsilon = \tilde{O}\left(1/\epsilon^2\right)$ consecutive samples. This is sufficient to detect a change and the total cost without a change is $p_\epsilon T n_\epsilon \Delta = \tilde{O}\left((\Delta/\epsilon)\sqrt{\ell T/K}\right)$. When an exponential schedule for the size of the changes $\epsilon = \Delta, 2\Delta, 4\Delta, \ldots$ is used, summing $(\Delta/\epsilon)\sqrt{\ell T/K}$ over the various ϵ and over the inferior arms gives a total regret contribution of $\tilde{O}\left(\sqrt{K\ell T}\right)$.

The number of steps until a change of size ϵ is detected is roughly $1/p_\epsilon$ such that the regret contribution is $\epsilon/p_\epsilon = \sqrt{KT/\ell}$. Summing over all L changes gives a total contribution of $\sqrt{KLT}$.

While in the formal analysis several other cases need to be considered —for example there could be a change even within a consecutive sample— the main idea of our algorithm is to use consecutive sampling as described above, choosing the right probability p_ϵ to start a consecutive sample of the right length n_ϵ.

Algorithm 1 ADSWITCH

Input: Time horizon T.
Initialization $\ell \leftarrow 0$, $t \leftarrow 0$.
Start a new episode:
  $\ell \leftarrow \ell + 1$.
  Set start of the episode $t_\ell \leftarrow t + 1$.
  $\mathrm{GOOD}_{t+1} = \{1, \dots, K\}$, $\mathrm{BAD}_{t+1} = \{\}$.
  Next time step:
    $t \leftarrow t + 1$.
    Add checks for bad arms:
      For all $a \in \mathrm{BAD}_t$, and all $i \geq 1$ with $2^{-i} \geq \tilde{\Delta}_\ell(a)/16$,
        with probability $2^{-i}\sqrt{\ell/(KT\log T)}$ add $\mathcal{S}_t(a) \leftarrow \mathcal{S}_t(a) \cup (2^{-i}, \lceil 2^{2i+1}\log T\rceil, t)$.
    Select an arm:
      Select $a_t = \arg\min_a\{\tau : a \notin \{a_\tau, \dots, a_{t-1}\}, a \in \mathrm{GOOD}_t \vee \mathcal{S}_t(a) \neq \{\}\}$.
      Receive reward r_t.
    Check for changes of good arms:
      If there is $a \in \mathrm{GOOD}_t$ and $t_\ell \leq s_1 \leq s_2 \leq t$ and $t_\ell \leq s \leq t$ such that condition (3)
        holds, then start a new episode.
    Check for changes of bad arms:
      If there is $a \in \mathrm{BAD}_t$ and $t_\ell \leq s \leq t$ such that condition (4) holds,
        then start a new episode.
      For $a \in \mathrm{BAD}_t$, $\mathcal{S}_{t+1}(a) \leftarrow \{(\epsilon, n, s) \in \mathcal{S}_t(a) : n_{[s,t]} < n\}$.
    Evict arms from GOOD_t***:***
      $\mathrm{BAD}_{t+1} = \mathrm{BAD}_t \cup \{a \in \mathrm{GOOD}_t | \exists s \geq t_\ell$ for which (1) holds$\}$.
      For evicted arms $a \in \mathrm{BAD}_{t+1} \setminus \mathrm{BAD}_t$, calculate $\tilde{\mu}_\ell(a)$ and $\tilde{\Delta}_\ell(a)$ according to (2), and
        set $\mathcal{S}_{t+1}(a) \leftarrow \{\}$.
      $\mathrm{GOOD}_{t+1} = \{1, \dots, K\} \setminus \mathrm{BAD}_{t+1}$.
    Continue with the next time step.

Remark 2 *Another approach for dealing with an unknown number of changes L is to run several copies of a bandit algorithm (for example EXP3.S) with different tunings, using a master bandit algorithm to manage these copies. While typically using a bandit algorithm on top of other bandit algorithms is problematic, the approach of Cheung et al. (2019) can be used to achieve the regret bound $\tilde{O}\left(\sqrt{KT\max\{L, T^{1/2}\}}\right)$ for unknown L even in the adversarial setting (Luo, 2019). This regret bound is optimal for large L but sub-optimal for small L. It is an open problem if optimal regret can be achieved in the adversarial setting without knowing L.*

3. The adaptive switching algorithm ADSWITCH

In this section we describe our algorithm ADSWITCH (depicted as Algorithm 1) for an unknown amount of changes. In order to achieve a regret bound that depends on the actual amount of change in the mean rewards, the algorithm needs to be able to detect (most of) these changes. At the same time, the algorithm cannot probe inferior arms too often without suffering large regret.

Our algorithm is essentially an elimination algorithm with restarts and proceeds in episodes $\ell = 1, 2, \ldots$. A new episode starts when the algorithm detects a change in the mean rewards. In each episode the arms are partitioned into GOOD and BAD arms. The *good* arms are those that are (so far) statistically indistinguishable from the optimal arm, and the *bad* arms are those that appear significantly worse than the optimal arm.

At the start of an episode all arms are good. An arm a is evicted from GOOD at time t, if there is sufficient evidence for its suboptimality through the condition[3]

$$\max_{a' \in \mathrm{GOOD}_t} \hat{\mu}_{[s,t]}(a') - \hat{\mu}_{[s,t]}(a) > \sqrt{\frac{C_1 \log T}{n_{[s,t]}(a) - 1}} \tag{1}$$

remove one which is the largest away from the terrible expr

for some s in the current episode with $n_{[s,t]} \geq 2$. Here $\hat{\mu}_{[s,t]}(a)$ denotes the observed mean reward for arm a during the time interval $[s, t]$, and $n_{[s,t]}(a)$ is the number of times arm a has been selected during this interval,

$$n_{[s,t]}(a) = \#\{s \leq \tau \leq t : a_\tau = a\}, \quad \hat{\mu}_{[s,t]}(a) = \frac{1}{n_{[s,t]}(a)} \sum_{\tau : s \leq \tau \leq t, a_\tau = a} r_t.$$

For a suitable constant C_1, condition (1) is a standard confidence bound on the mean rewards.

When an arm a is evicted from the good arms in episode ℓ, then its observed mean reward and the gap to the arm a' that caused the eviction, are recorded,

$$\tilde{\mu}_\ell(a) \leftarrow \hat{\mu}_{[s,t]}(a), \quad \tilde{\Delta}_\ell(a) \leftarrow \max_{a' \in \mathrm{GOOD}_t} \hat{\mu}_{[s,t]}(a') - \hat{\mu}_{[s,t]}(a). \tag{2}$$

These quantities will be used to check for changes in the mean reward of the evicted arm.

To check at time t, if a good arm a has changed, we use a condition similar to (1),

$$\left|\hat{\mu}_{[s_1,s_2]}(a) - \hat{\mu}_{[s,t]}(a)\right| > \sqrt{\frac{2 \log T}{n_{[s_1,s_2]}(a)}} + \sqrt{\frac{2 \log T}{n_{[s,t]}(a)}} \tag{3}$$

for some $s_1 \leq s_2$ and s within the current episode.

To check for changes of bad arms is more complicated, since these arms can be selected only rarely without causing large regret. These checks are done by a variant of *consecutive sampling* as described in Section 2.1.3. We associate each bad arm a with a set $\mathcal{S}_t(a)$ of sampling obligations (ϵ, n, s) as follows. Each $\mathcal{S}_t(a) \subset \mathbb{R} \times \mathbb{N} \times \mathbb{N}$ is a set of triples (ϵ, n, s), where $\epsilon = 2^{-i}$, $i \geq 1$, is the magnitude of change the algorithm seeks to detect, $n = \lceil 2(\log T)/\epsilon^2 \rceil$ is the number of samples needed for a statistically significant test, and s is the time when the collection of samples has started. After having received n rewards from arm a, the sampling obligation is removed from $\mathcal{S}_t(a)$. These sampling obligations cause the algorithm to select a bad arm that otherwise would not be selected. The test for a change is similar to the check for the good arms: a new episode is started, if for some s in the current episode,

$$\left|\hat{\mu}_{[s,t]}(a) - \tilde{\mu}_\ell(a)\right| > \tilde{\Delta}_\ell(a)/4 + \sqrt{\frac{2 \log T}{n_{[s,t]}(a)}}, \tag{4}$$

3. The constant C_1 in condition (1) needs to be sufficiently large. A suitable value can be derived from the regret analysis.

comparing the current mean reward with the mean reward when the arm was evicted from the set of good arms.

To ensure the right amount of checking, at any time t and for any $\epsilon = 2^{-i} \geq \tilde{\Delta}_\ell(a)/16$, the sampling obligation (ϵ, n, t) is added to $\mathcal{S}_t(a)$ with probability $\epsilon\sqrt{\ell/(KT\log T)}$. The intuition behind the choice of this probability is given in Section 2.1.3.

Finally, at time t the algorithm selects the arm a_t that has been selected least recently among the good arms and those bad arms with non-empty sampling obligations $\mathcal{S}_t(a)$. This selects the good arms in a round robin fashion and also ensures that almost consecutive samples are generated for the bad arms that need to be checked.

Remark 3 *This version of our algorithm is not optimized for runtime but for simpler arguments in the regret analysis. The most expensive step is the check for changes of the good arms, condition (3), with runtime $O\left(Kt^3\right)$ in time step t. This time complexity can be significantly reduced, if not all intervals $[s_1, s_2]$ and $[s, t]$ are checked, but only intervals of certain lengths, say $2^k \log T$ for $k = 3, 4, \ldots$ By storing for these lengths the maximal and minimal values of $\hat{\mu}_{[s_1,s_2]}(a)$ so far, the time complexity can be reduced to $O\left(K(\log T)^2\right)$ per time step. The checks of conditions (1) and (4) can be treated similarly.*

4. Regret analysis

4.1. Preliminaries

Lemma 4 (Azuma-Hoeffding inequality) *For a martingale difference sequence $X_1, \ldots, X_n$ with support of size 1 for all X_i,*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_i\right| \geq \gamma n\right\} \leq 2\exp\{-2\gamma^2 n\}.$$

We denote the number of changes in time interval $[s, t]$ by

$$L[s, t] = \#\{s < \tau \leq t | \exists a : \mu_{\tau-1}(a) \neq \mu_\tau(a)\}.$$

Lemma 5 *With probability $1 - 2K/T^2$, for all $1 \leq s \leq t \leq T$ with $L[s, t] = 0$, and all arms a,*

$$\left|\hat{\mu}_{[s,t]}(a) - \mu_s(a)\right| < \sqrt{\frac{2\log T}{n_{[s,t]}(a)}}.$$

Proof We fix time s and arm a. Let $s \leq \tau_1 < \tau_2 < \cdots$ be the times when arm a is selected, $a_{\tau_i} = a$. Then $X_i = r_{\tau_i} - \mu_{\tau_i}(a)$ are martingale differences and by Lemma 4 we get $\mathbb{P}\left\{|\sum_{i=1}^{n} X_i| \geq \sqrt{2n\log T}\right\} \leq 2T^{-4}$. By a union bound we get $\mathbb{P}\left\{\exists n : n \leq T : |\sum_{i=1}^{n} X_i| \geq \sqrt{2n\log T}\right\} \leq 2T^{-3}$. Thus with probability $1 - 2T^{-3}$ we have for all t,

$$\left|\hat{\mu}_{[s,t]}(a) - \mu_s(a)\right| = \left|\frac{1}{n_{[s,t]}(a)} \sum_{i=1}^{n_{[s,t]}} [r_{\tau_i} - \mu_s(a)]\right| = \left|\frac{1}{n_{[s,t]}(a)} \sum_{i=1}^{n_{[s,t]}} X_i\right| < \sqrt{\frac{2\log T}{n_{[s,t]}(a)}}.$$

A union bound over s and a completes the proof. ■

Since the error probability $2K/T^2$ in Lemma 5 causes only diminishing regret, we assume in all the following, that all inequalities of the lemma are satisfied.

Assumption 6 *For all $1 \le s \le t \le T$ with no change between s and t, and all arms a,*

$$\left|\hat{\mu}_{[s,t]}(a) - \mu_s(a)\right| < \sqrt{\frac{2\log T}{n_{[s,t]}(a)}}.$$

4.2. Counting the number of episodes

The next lemma shows that the algorithm starts a new episode only if there is a change in the current episode. We denote by $\tilde{L}$ the total number of episodes. The last episode ends at time T, and for notational convenience we define the start time of the next episode by $t_{\tilde{L}+1} = T+1$.

Lemma 7 *If Assumption 6 holds, then for all episodes $\ell < \tilde{L}$, $L[t_\ell, t_{\ell+1}-1] > 0$.*

Proof The proof is by contradiction, assuming that episode ℓ is terminated at time $t > t_\ell$ but $L[t_\ell, t] = 0$. We first consider the start of a new episode when condition (3) is met. Then there is an arm a and times $t_\ell \le s_1 \le s_2 \le t$ and $t_\ell \le s \le t$ with

$$\left|\hat{\mu}_{[s_1,s_2]}(a) - \hat{\mu}_{[s,t]}(a)\right| > \sqrt{\frac{2\log T}{n_{[s_1,s_2]}(a)}} + \sqrt{\frac{2\log T}{n_{[s,t]}(a)}}.$$

Since $L[t_\ell, t] = 0$ and $\mu_{s_1}(a) = \mu_s(a)$, this contradicts Assumption 6.

Now we consider the start of a new episode when condition (4) is met. Then there is an arm $a \in \mathrm{BAD}_t$ and a time $t_\ell \le s \le t$ with

$$\left|\hat{\mu}_{[s,t]}(a) - \tilde{\mu}_\ell(a)\right| > \tilde{\Delta}_\ell(a)/4 + \sqrt{\frac{2\log T}{n_{[s,t]}(a)}}.$$

Let $[s', t']$ be the time interval on which the eviction of arm a from the good arms was based in (1). Then $\tilde{\mu}_\ell(a) = \hat{\mu}_{[s',t']}(a)$ and

$$\tilde{\Delta}_\ell(a) > \sqrt{\frac{C_1 \log T}{n_{[s',t']} - 1}} > 4\sqrt{\frac{2\log T}{n_{[s',t']}}} \tag{5}$$

for sufficiently large C_1. Together with the above inequality, this contradicts Assumption 6. ■

From Lemma 7 we get immediately that the total number of episodes is bounded by the number of changes, $\tilde{L} \le L$.

4.3. Properties of the selection rule

An arm a is eligible at time t, if $a \in \mathrm{GOOD}_t$ or $a \in \mathrm{BAD}_t$ and $\mathcal{S}_t(a) \neq \{\}$. Since the algorithm selects the arm that has been selected least recently among the eligible arms, all eligible arms are selected almost equally often, the maximal difference being 1. In particular, if arm a has been eligible throughout an interval $[s, t]$, then for any arm a',

$$n_{[s,t]}(a) \ge n_{[s,t]}(a') - 1.$$

Furthermore, if an arm a is eligible throughout the interval $[s, t]$, then

$$n_{[s,t]}(a) \ge \lfloor (t-s+1)/K \rfloor.$$

4.4. Dividing episodes into intervals with no change

In the regret analysis we will rely on time intervals without change. Thus for each episode ℓ we consider all change points $\beta_{\ell,1}, \ldots, \beta_{\ell,m_\ell}$ in episode ℓ, $L[t_\ell, t_{\ell+1} - 1] = m_\ell$, with

$$t_\ell = \beta_{\ell,0} \leq \beta_{\ell,1} < \ldots < \beta_{\ell,m_\ell} < \beta_{\ell,m_\ell+1} = t_{\ell+1}$$

and $L[\beta_{\ell,i}, \beta_{\ell,i+1} - 1] = 0$ for $i = 0, \ldots, m_\ell$ and $L[\beta_{\ell,i} - 1, \beta_{\ell,i}] = 1$ for $i = 1, \ldots, m_\ell$. We denote the intervals with no change by

$$I_{\ell,i} = [\beta_{\ell,i}, \beta_{\ell,i+1} - 1]$$

for $i = 0, \ldots, m_\ell$. Since each episode is split into at most $m_\ell + 1$ intervals, over all episodes there are at most $L + \tilde{L} \leq 2L$ such intervals.

4.5. Distinguishing the sources of regret

In this section we decompose the regret $\sum_{t=1}^{T} [\max_a \mu_t(a) - \mu_t(a_t)]$ horizontally and vertically. By horizontally we mean the decomposition of the regret

$$\max_a \mu_t(a) - \mu_t(a_t) = [\max_{a \in \mathrm{GOOD}_t} \mu_t(a) - \mu_t(a_t)] + [\max_a \mu_t(a) - \max_{a \in \mathrm{GOOD}_t} \mu_t(a)]$$

into the regret in respect to the best good arm, and the regret of the best good arm in respect to the optimal arm. We denote by

$$a_t^* = \arg\max_a \mu_t(a)$$

the optimal arm and by

$$a_t^g = \arg \max_{a \in \mathrm{GOOD}_t} \mu_t(a)$$

the best good arm at time t. (If there are several optimal arms, then an arbitrary one can be chosen.)

By vertical decomposition we mean the classification of the time steps into several classes, depending on the selections of the algorithm. For the regret in respect to the best good arm, we are distinguishing the following cases for each episode ℓ:

1. A good arm is selected by the algorithm,

   $$\mathcal{G}_{\ell,1} = \{t_\ell \leq t < t_{\ell+1} : a_t \in \mathrm{GOOD}_t\}.$$

   The regret in this case is similar to the regret in the stationary bandit problem, when the best arm has not been distinguished yet.

2. A bad arm is selected, and its regret is not much larger than its eviction gap $\tilde{\Delta}_\ell$,

   $$\mathcal{G}_{\ell,2} = \{t_\ell \leq t < t_{\ell+1} : a_t \in \mathrm{BAD}_t, \mu_t(a_t^g) - \mu_t(a_t) \leq 4\tilde{\Delta}_\ell(a_t)\}.$$

   The regret in this case is for the effort of checking whether a previously bad arm has become optimal.

3. A bad arm is selected, its regret is large, and the mean reward is far from the mean reward when it was evicted,

$$\mathcal{G}_{\ell,3} = \{t_\ell \le t < t_{\ell+1} : a_t \in \mathrm{BAD}_t, \mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\ell(a_t), \\ \tilde{\mu}_\ell(a_t) - \mu_t(a_t) > (\mu_t(a_t^g) - \mu_t(a_t))/2\}.$$

   In this case the mean reward of the bad arm has decreased significantly, causing larger regret when this arm is selected.

4. A bad arm is selected, its regret is large, but the mean reward is relatively close to the mean reward when it was evicted,

$$\mathcal{G}_{\ell,4} = \{t_\ell \le t < t_{\ell+1} : a_t \in \mathrm{BAD}_t, \mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\ell(a_t), \\ \tilde{\mu}_\ell(a_t) - \mu_t(a_t) \le (\mu_t(a_t^g) - \mu_t(a_t))/2\}.$$

   In this case the best good arm has significantly improved (compared to the time when a_t was evicted), and additional regret is caused by the resulting larger gap between the best good and the bad arm.

For the regret of the best good arm in respect to the optimal arm, we distinguish two cases. Obviously there is no such regret if the optimal arm is among the good arms.

1. The optimal arm is among the bad arms and its mean reward is close to the mean reward when it was evicted,

$$\mathcal{B}_{\ell,1} = \{t_\ell \le t < t_{\ell+1} : a_t^* \in \mathrm{BAD}_t, \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) \le \tilde{\Delta}_\ell(a_t^*)/2\}.$$

   In this case the mean rewards of the good arms have significantly decreased, causing regret when the algorithm keeps selecting them.

2. The optimal arm is among the bad arms and its mean reward is far from the mean reward when it was evicted,

$$\mathcal{B}_{\ell,2} = \{t_\ell \le t < t_{\ell+1} : a_t^* \in \mathrm{BAD}_t, \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) > \tilde{\Delta}_\ell(a_t^*)/2\}.$$

   In this case the mean reward of the currently optimal arm may have significantly improved, causing regret when not selected.

In the following sections we show that in each of these cases the respective regret is of order $\sqrt{KLT\log T}$. Summing over all these cases gives the result of Theorem 1.

Because of space constraints we give only brief intuition about the cases $t \in \mathcal{B}_{\ell,1}$ and $t \in \mathcal{B}_{\ell,2}$. Full proofs are provided in the appendix.

4.6. Bounding the regret in respect to a_t^g

4.6.1. CASE $t \in \mathcal{G}_{\ell,1}$

In this case $a_t \in \mathrm{GOOD}_t$. Let $t \in I_{\ell,i}$ for some interval without change $I_{\ell,i} = [\beta_{\ell,i}, \beta_{\ell,i+1} - 1]$ in episode ℓ. Since $a_t \in \mathrm{GOOD}_t$, a_t was not evicted at time $t-1$, and by (1) we have

$$\hat{\mu}_{[\beta_{\ell,i},t-1]}(a_t^g) - \hat{\mu}_{[\beta_{\ell,i},t-1]}(a_t) \le \sqrt{\frac{C_1 \log T}{n_{[\beta_{\ell,i},t-1]}(a_t) - 1}}$$

if $n_{[\beta_{\ell,i},t-1]}(a_t) \geq 2$. By Assumption 6 we get

$$\begin{aligned}\mu_t(a_t^g) - \mu_t(a_t) &\leq \hat{\mu}_{[\beta_{\ell,i},t-1]}(a_t^g) - \hat{\mu}_{[\beta_{\ell,i},t-1]}(a_t) + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t-1]}(a_t^g)}} + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t-1]}(a_t)}} \\ &\leq \sqrt{\frac{C_1 \log T}{n_{[\beta_{\ell,i},t-1]}(a_t) - 1}} + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t-1]}(a_t^g)}} + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t-1]}(a_t)}} \\ &\leq (\sqrt{C_1} + 2\sqrt{2})\sqrt{\frac{\log T}{n_{[\beta_{\ell,i},t-1]}(a_t) - 1}},\end{aligned}$$

since a_t^g and a_t are both good arms and thus $n_{[\beta_{\ell,i},t-1]}(a_t^g) \geq n_{[\beta_{\ell,i},t-1]}(a_t) - 1$ (see Section 4.3). Summing over $t \in I_{\ell,i} \cap \mathcal{G}_{\ell,1}$ with $a_t = a$ for some arm a gives

$$\begin{aligned}\sum_{t \in I_{\ell,i} \cap \mathcal{G}_{\ell,1}, a_t = a} [\mu_t(a_t^g) - \mu_t(a_t)] &\leq (\sqrt{C_1} + 2\sqrt{2})\sqrt{\log T}\left(2 + \sum_{k=2}^{n_{[I_{\ell,i}]}(a)} \sqrt{\frac{1}{k-1}}\right) \\ &\leq (\sqrt{C_1} + 2\sqrt{2})\sqrt{\log T}\left(2 + 2\sqrt{n_{[I_{\ell,i}]}(a)}\right).\end{aligned}$$

Since

$$\sum_{\ell,i,a} n_{[I_{\ell,i}]}(a) = T$$

and there are at most $2L$ intervals $I_{\ell,i}$, summing over all intervals and all arms gives

$$\sum_{\ell=1}^{\tilde{L}} \sum_{t \in \mathcal{G}_{\ell,1}} [\mu_t(a_t^g) - \mu_t(a_t)] = O\left(\sqrt{KLT\log T}\right). \tag{6}$$

4.6.2. CASE $t \in \mathcal{G}_{\ell,2}$

In this case $a_t \in \mathrm{BAD}_t$ and $\mu_t(a_t^g) - \mu_t(a_t) \leq 4\tilde{\Delta}_\ell(a_t)$. We bound the expected number of times that such a bad arm a is selected.

A bad arm a is selected at time t, $a_t = a$, only if there is $(\epsilon, n, s) \in \mathcal{S}_t(a)$ with $s \leq t$ and $n_{[s,t]}(a) < n$. We recall that $\epsilon = 2^{-i} \geq \tilde{\Delta}_\ell(a)/16$ for some $i \geq 1$, and $n = \lceil 2(\log T)/\epsilon^2 \rceil$. Thus for $t \in \mathcal{G}_{\ell,2}$,

$$\mu_t(a_t^g) - \mu_t(a_t) \leq 64\epsilon.$$

Furthermore, (ϵ, n, s) is added to $\mathcal{S}_s(a_t)$ with probability $\epsilon\sqrt{\ell/(KT\log T)}$ if $a \in \mathrm{BAD}_s$. Thus the expected number of times that a specific bad arm a is selected in episode ℓ for a specific ϵ, is at most

$$\epsilon(t_{\ell+1} - t_\ell)\sqrt{\frac{\ell}{KT\log T}}(2(\log T)/\epsilon^2 + 1).$$

Summing over the possible $\epsilon = 2^{-i}$ with $\epsilon \geq [\mu_t(a_t^g) - \mu_t(a)]/64$ gives

$$(t_{\ell+1} - t_\ell)\sqrt{\frac{\ell}{KT\log T}}\left(\frac{256\log T}{\mu_t(a_t^g) - \mu_t(a)} + 1\right)$$

as an upper bound on the expected number of times arm a is selected in $\mathcal{G}_{\ell,2}$. Summing over all arms we get for the respective expected regret

$$\sum_{t\in\mathcal{G}_{\ell,2}} [\mu_t(a_t^g) - \mu_t(a_t)] \leq K(t_{\ell+1} - t_\ell)\sqrt{\frac{\ell}{KT\log T}}\,(256\log T + 1)\,,$$

and summing over all episodes gives

$$\sum_{\ell}\sum_{t\in\mathcal{G}_{\ell,2}} [\mu_t(a_t^g) - \mu_t(a_t)] = O\left(\sqrt{KLT\log T}\right). \tag{7}$$

4.6.3. CASE $t \in \mathcal{G}_{\ell,3}$

In this case $a_t \in \mathrm{BAD}_t$, $\mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\ell(a_t)$, and $\tilde{\mu}_\ell(a_t) - \mu_t(a_t) > [\mu_t(a_t^g) - \mu_t(a_t)]/2$. Thus the mean reward for arm a_t is significantly worse than $\tilde{\mu}_\ell(a_t)$. We show that with this condition, arm a_t can be selected only a few times in an interval $I_{\ell,i}$. This is because otherwise the algorithm would detect the change.

4.6.4. CASE $t \in \mathcal{G}_{\ell,4}$

In this case $a_t \in \mathrm{BAD}_t$, $\mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\ell(a_t)$, and $\tilde{\mu}_\ell(a_t) - \mu_t(a_t) \leq (\mu_t(a_t^g) - \mu_t(a_t))/2$. The major part of the regret contribution comes from the increase of the mean of a_t^g, the best good arm. But since this arm is selected often, the change of the reward can be detected quickly by condition (3), obtaining the desired regret bound.

4.7. Bounding the regret of $\mu_t(a_t^g)$ in respect to $\mu_t(a_t^*)$

4.7.1. CASE $t \in \mathcal{B}_{\ell,1}$

In this case $a_t^* \in \mathrm{BAD}_t$ and $\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) \leq \tilde{\Delta}_\ell(a_t^*)/2$. This means that the reward of the bad arm a_t^* has not changed much, but that the rewards of the good arms have decreased. There are basically two cases: (a) The arm a' that evicted a_t^* from the good arms is still a good arm. Then the reward of this arm has changed significantly which is detected by condition (3) of the algorithm. (b) Arm a' has also been evicted from the good arms. This case is more complicated, but we can show that also in this case the best good arm a_t^* has changed significantly.

4.7.2. CASE $t \in \mathcal{B}_{\ell,2}$

In this case a_t^* is among the bad arms and $\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) > \tilde{\Delta}_\ell(a_t^*)/2$. Since the reward of the best good arm $\mu_t(a_t^g)$ cannot be much below $\tilde{\mu}_\ell(a_t^*)$ without causing the start of a new episode, $\mu_t(a_t^*) - \mu_t(a_t^g) \lesssim \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)$.

If $\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)$ is small, then the contribution to the regret is also small and is easily dealt with. If $\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)$ is large, though, then the larger $\mu_t(a_t^*)$ needs to be detected quickly. The algorithm needs roughly $n = (\log T)/\epsilon^2$ samples, $\epsilon \approx \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)$, to detect the change. These samples are provided by a sampling obligation (ϵ, n, s). By the definition of the algorithm, such a sampling obligation is added with probability $\epsilon\sqrt{\ell/(KT\log T)}$. Thus it takes roughly time $(1/\epsilon)\sqrt{KT(\log T)/\ell}$ until the sampling obligation is added and the change is detected, causing $\sqrt{KT(\log T)/\ell}$ regret. Summing over the episodes gives the desired bound $\sqrt{KLT(\log T)}$.

The actual proof is a bit more complicated, because arms may change during the sampling.

5. Conclusion

Extending the work in (Auer et al., 2018), we have constructed the first algorithm for the stochastic multi-armed bandit problem with abrupt changes of the reward distributions that achieves optimal regret bounds without knowing the number of changes in advance. The main technical contribution is the delicate testing schedule of the apparently inferior arms. This testing is necessary to detect when a previously inferior arm becomes the best arm.

We note that our algorithm (without any change) also provides optimal regret bounds in terms of total variation. These optimal bounds have also been achieved in (Chen et al., 2019), which provides also optimal bounds for the more general stochastic contextual bandits setting.

Regarding the adversarial bandit setting, it remains an open problem to construct an algorithm with optimal regret bounds without a priori tuning in respect to the number of arm changes.

Acknowledgments

This work has been supported by the Austrian Science Fund (FWF): I 3437-N33 in the framework of the CHIST-ERA ERA-NET (DELTA project).

References

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283, 2017.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 116–120, 2016.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best arm with an unknown number of distribution changes. In *14th European Workshop on Reinforcement Learning, EWRL 2018*, 2018.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems 27, NIPS 2014*, pages 199–207, 2014.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *CoRR*, abs/1803.06971, 2018. URL http://arxiv.org/abs/1803.06971.

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

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *COLT 2012 - The 25th Annual Conference on Learning Theory*, pages 42.1–42.23, 2012.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *32nd Annual Conference on Learning Theory (COLT)*, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, pages 1079–1087, 2019.

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

Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michèle Sebag. Multi-armed bandit, dynamic environments and meta-bandits. *NIPS-2006 workshop, Online trading between exploration and exploitation*, 2006.

Levente Kocsis and Csaba Szepesvári. Discounted UCB. *2nd PASCAL Challenges Workshop*, 2006.

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

Haipeng Luo. Personal communication, 2019.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory, COLT 2018*, pages 1739–1776, 2018.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless Markov bandits. *Theoretical Computer Science*, 558:62–76, 2014.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning, ICML 2014*, pages 1287–1295, 2014.

Alex Slivkins and Eli Upfal. Adapting to a changing environment: The Brownian restless bandits. In *21st Conference on Learning Theory, COLT 2008*, pages 343–354, 2008.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pages 1177–1184, 2009.

Appendix A. Some proof details of Section 4.6
Bounding the regret in respect to (a_t^g)

A.1. Case $t \in \mathcal{G}_{\ell,3}$

In this case $a_t \in \text{BAD}_t$, $\mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\ell(a_t)$, and $\tilde{\mu}_\ell(a_t) - \mu_t(a_t) > [\mu_t(a_t^g) - \mu_t(a_t)]/2$. Thus the mean reward for arm a_t is significantly worse than $\tilde{\mu}_\ell(a_t)$. We show that with this condition, arm a_t can be selected only a few times in an interval $I_{\ell,i}$. This is because otherwise the algorithm would detect the change.

If condition (4) is not triggered, we have for bad arm a_t and $t < t_{\ell+1} - 1$ that

$$\tilde{\mu}_\ell(a_t) - \hat{\mu}_{[t_\ell,t]}(a_t) \le \tilde{\Delta}_\ell(a_t)/4 + \sqrt{\frac{2\log T}{n_{[t_\ell,t]}(a_t)}}.$$

Using the condition of $\mathcal{G}_{\ell,3}$ and Assumption 6, we get

$$\begin{aligned}\mu_t(a_t^g) - \mu_t(a_t) &< 2[\tilde{\mu}_\ell(a_t) - \mu_t(a_t)] < 2\left[\tilde{\mu}_\ell(a_t) - \hat{\mu}_{[t_\ell,t]}(a_t) + \sqrt{\frac{2\log T}{n_{[t_\ell,t]}(a_t)}}\right] \\ &\le 2\left[\tilde{\Delta}_\ell(a_t)/4 + 2\sqrt{\frac{2\log T}{n_{[t_\ell,t]}(a_t)}}\right] < [\mu_t(a_t^g) - \mu_t(a_t)]/8 + 4\sqrt{\frac{2\log T}{n_{[t_\ell,t]}(a_t)}}\end{aligned}$$

and by solving for $\mu_t(a_t^g) - \mu_t(a_t)$,

$$\mu_t(a_t^g) - \mu_t(a_t) < \frac{32}{7}\sqrt{\frac{2\log T}{n_{[t_\ell,t]}(a_t)}}.$$

Summing over all arms and episodes gives

$$\sum_\ell \sum_{t\in\mathcal{G}_{\ell,3}} [\mu_t(a_t^g) - \mu_t(a_t)] = O\left(\sqrt{KLT\log T}\right). \tag{8}$$

A.2. Case $t \in \mathcal{G}_{\ell,4}$

In this case $a_t \in \text{BAD}_t$, $\mu_t(a_t^g) - \mu_t(a_t) > 4\tilde{\Delta}_\ell(a_t)$, and $\tilde{\mu}_\ell(a_t) - \mu_t(a_t) \le (\mu_t(a_t^g) - \mu_t(a_t))/2$. We bound the number of times that a bad arm a is selected within an interval without change $I_{\ell,i} = [\beta_{\ell,i}, \beta_{\ell,i+1} - 1]$, while the above condition holds. Let $[s', t']$ be the interval and a' be the arm that caused the eviction of a_t from the good arms by condition (1),

$$\hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t) > \sqrt{\frac{C_1\log T}{n_{[s',t']}(a_t) - 1}},$$

with $\tilde{\mu}_\ell(a_t) = \hat{\mu}_{[s',t']}(a_t)$ and $\tilde{\Delta}_\ell(a_t) = \hat{\mu}_{[s',t']}(a') - \tilde{\mu}_\ell(a_t)$. By the construction of the algorithm (see Section 4.3), $n_{[s',t']}(a_t) \le n_{[s',t']}(a') + 1$ and $n_{[\beta_{\ell,i},t]}(a_t) \le n_{[\beta_{\ell,i},t]}(a_t^g) + 1 \le 2n_{[\beta_{\ell,i},t]}(a_t^g)$.

By Assumption 6,

$$\begin{aligned}\hat{\mu}_{[\beta_{\ell,i},t]}(a_t^g) - \hat{\mu}_{[s',t']}(a_t^g) &> \mu_t(a_t^g) - \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} - \hat{\mu}_{[s',t']}(a') \\ &= \mu_t(a_t^g) - \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} - \tilde{\mu}_\ell(a_t) - \tilde{\Delta}_\ell(a_t) \\ &= [\mu_t(a_t^g) - \mu_t(a_t)] - \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} - [\tilde{\mu}_\ell(a_t) - \mu_t(a_t)] - \tilde{\Delta}_\ell(a_t) \\ &\geq \frac{1}{4}[\mu_t(a_t^g) - \mu_t(a_t)] - \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}}.\end{aligned}$$

For $t < t_{\ell+1} - 1$, we get from condition (3) that

$$\begin{aligned}\hat{\mu}_{[\beta_{\ell,i},t]}(a_t^g) - \hat{\mu}_{[s',t']}(a_t^g) &\leq \sqrt{\frac{2\log T}{n_{[s',t']}(a_t^g)}} + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \\ &\leq \sqrt{\frac{2\log T}{n_{[s',t']}(a_t) - 1}} + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \\ &\leq \sqrt{\frac{2}{C_1}}\tilde{\Delta}_\ell(a_t) + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \\ &\leq \frac{1}{4}\sqrt{\frac{2}{C_1}}[\mu_t(a_t^g) - \mu_t(a_t)] + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}}.\end{aligned}$$

Putting the lower and the upper bound together we find

$$\frac{1}{4}\left(1 - \sqrt{\frac{2}{C_1}}\right)[\mu_t(a_t^g) - \mu_t(a_t)] \leq 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \leq 2\sqrt{\frac{4\log T}{n_{[\beta_{\ell,i},t]}(a_t)}}.$$

Thus the contribution to the regret for some arm a during $I_{\ell,i}$ in respect to $\mathcal{G}_{\ell,4}$ is at most

$$32\sqrt{(\log T) n_{[I_{\ell,i}]}(a)} \Big/ \left(1 - \sqrt{\frac{2}{C_1}}\right).$$

Summing over all intervals $I_{\ell,i}$ and all arms a, we get

$$\sum_\ell \sum_{t \in \mathcal{G}_{\ell,4}} [\mu_t(a_t^g) - \mu_t(a_t)] = O\left(\sqrt{KLT\log T}\right) \tag{9}$$

Appendix B. Proof details of Section 4.7
Bounding the regret of $\mu_t(a_t^g)$ in respect to $\mu_t(a_t^*)$

B.1. Case $t \in \mathcal{B}_{\ell,1}$

In this case $a_t^* \in \mathrm{BAD}_t$ and $\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) \leq \tilde{\Delta}_\ell(a_t^*)/2$. This implies that the mean of the best good arm has dropped significantly. Thus there cannot be too many such steps without detecting the change.

We consider some interval without change $I_{\ell,i} = [\beta_{\ell,i}, \beta_{\ell,i+1} - 1]$, $t \in I_{\ell,i}$. Let $[s', t']$ be the time interval and a' the arm that caused the eviction of a_t^* from the good arms by condition (1),

$$\hat{\mu}_{[s',t']}(a') - \hat{\mu}_{[s',t']}(a_t^*) > \sqrt{\frac{C_1 \log T}{n_{[s',t']}(a_t^*) - 1}}.$$

If a' is not a good arm anymore, $a' \notin \mathrm{GOOD}_t$, then its eviction was caused by some interval $[s_1, s_2]$ and some arm a'' with

$$\hat{\mu}_{[s_1,s_2]}(a'') - \hat{\mu}_{[s_1,s_2]}(a') > \sqrt{\frac{C_1 \log T}{n_{[s_1,s_2]}(a') - 1}}.$$

Since the episode has not stopped at time s_2, we have by (3),

$$\begin{aligned}
\sqrt{\frac{C_1 \log T}{n_{[s_1,s_2]}(a') - 1}} &< \hat{\mu}_{[s_1,s_2]}(a'') - \hat{\mu}_{[s_1,s_2]}(a') \\
&< \hat{\mu}_{[s',t']}(a'') - \hat{\mu}_{[s',t']}(a') \\
&\quad + \sqrt{\frac{2 \log T}{n_{[s_1,s_2]}(a'')}} + \sqrt{\frac{2 \log T}{n_{[s_1,s_2]}(a')}} + \sqrt{\frac{2 \log T}{n_{[s',t']}(a'')}} + \sqrt{\frac{2 \log T}{n_{[s',t']}(a')}} \\
&\leq \sqrt{\frac{2 \log T}{n_{[s_1,s_2]}(a'')}} + \sqrt{\frac{2 \log T}{n_{[s_1,s_2]}(a')}} + \sqrt{\frac{2 \log T}{n_{[s',t']}(a'')}} + \sqrt{\frac{2 \log T}{n_{[s',t']}(a')}} \\
&\leq 2\sqrt{\frac{2 \log T}{n_{[s_1,s_2]}(a') - 1}} + 2\sqrt{\frac{2 \log T}{n_{[s',t']}(a') - 1}} \\
&\leq \frac{1}{2}\sqrt{\frac{C_1 \log T}{n_{[s_1,s_2]}(a') - 1}} + 2\sqrt{\frac{2 \log T}{n_{[s',t']}(a') - 1}}
\end{aligned}$$

for sufficiently large C_1, and therefore

$$\sqrt{\frac{C_1 \log T}{n_{[s_1,s_2]}(a') - 1}} \leq 4\sqrt{\frac{2 \log T}{n_{[s',t']}(a') - 1}},$$

and

$$n_{[s',t']}(a') \leq \frac{64}{C_1} n_{[s_1,s_2]}(a').$$

Since a' was evicted based on interval $[s_1, s_2]$ and a_t^g was not, we have

$$\hat{\mu}_{[s_1,s_2]}(a_t^g) > \hat{\mu}_{[s_1,s_2]}(a') + \sqrt{\frac{C_1 \log T}{n_{[s_1,s_2]}(a') - 1}} - \sqrt{\frac{C_1 \log T}{n_{[s_1,s_2]}(a_t^g) - 1}}.$$

Then

$$\begin{aligned}\mu_t(a_t^g) &\geq \hat{\mu}_{[\beta_{\ell,i},t]}(a_t^g) - \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \geq \hat{\mu}_{[s_1,s_2]}(a_t^g) - 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} - \sqrt{2\frac{\log T}{n_{[s_1,s_2]}(a_t^g)}} \\ &\geq \hat{\mu}_{[s_1,s_2]}(a') - 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} - (\sqrt{2}+\sqrt{C_1})\sqrt{\frac{\log T}{n_{[s_1,s_2]}(a_t^g)-1}} \\ &\geq \hat{\mu}_{[s',t']}(a') - 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} - (\sqrt{2}+\sqrt{C_1})\sqrt{\frac{\log T}{n_{[s_1,s_2]}(a')-2}} \\ &\quad - \sqrt{\frac{2\log T}{n_{[s_1,s_2]}(a')}} - \sqrt{\frac{2\log T}{n_{[s',t']}(a')}} \\ &\geq \hat{\mu}_{[s',t']}(a') - 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)-1}} - \sqrt{\frac{C_2\log T}{n_{[s',t']}(a')}}\end{aligned}$$

for some suitable constant C_2 independent of C_1.

Also if a' is still among the good arms, $a' \in \mathrm{GOOD}_t$, we get for $t < t_{\ell+1}-1$, by Assumption 6 and the checking condition (3), that

$$\begin{aligned}\mu_t(a_t^g) \geq \mu_t(a') &\geq \hat{\mu}_{[\beta_{\ell,i},t]}(a') - \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a')}} \geq \hat{\mu}_{[s',t']}(a') - 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a')}} - \sqrt{\frac{2\log T}{n_{[s',t']}(a')}} \\ &\geq \hat{\mu}_{[s',t']}(a') - 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)-1}} - \sqrt{\frac{2\log T}{n_{[s',t]}(a')}} \\ &\geq \hat{\mu}_{[s',t']}(a') - 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)-1}} - \sqrt{\frac{C_2\log T}{n_{[s',t']}(a')}}.\end{aligned}$$

Since

$$\begin{aligned}\hat{\mu}_{[s',t']}(a') &= \tilde{\mu}_\ell(a_t^*) + \tilde{\Delta}_\ell(a_t^*) \geq \mu_t(a_t^*) + \tilde{\Delta}_\ell(a_t^*)/2 \geq \mu_t(a_t^*) + \frac{1}{2}\sqrt{\frac{C_1\log T}{n_{[s',t']}(a_t^*)-1}} \\ &\geq \mu_t(a_t^*) + \frac{1}{2}\sqrt{\frac{C_1\log T}{n_{[s',t']}(a')}},\end{aligned}$$

we get for sufficiently large C_1,

$$\mu_t(a_t^*) - \mu_t(a_t^g) \leq 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)-1}} \leq 2\sqrt{\frac{2\log T}{\lfloor (t-\beta_{\ell,i}+1)/K \rfloor - 1}}.$$

Summing over $t \in I_{\ell,i}$ gives

$$\sum_{t\in I_{\ell,i}\cap\mathcal{B}_{\ell,1}} [\mu_t(a_t^*) - \mu_t(a_t^g)] \leq 2K + 4\sqrt{2K|I_{\ell,i}|\log T}$$

and summing over all intervals gives

$$\sum_\ell \sum_{t\in\mathcal{B}_{\ell,1}} [\mu_t(a_t^*) - \mu_t(a_t^g)] = O\left(\sqrt{KLT\log T}\right). \tag{10}$$

B.2. Case $t \in \mathcal{B}_{\ell,2}$

In this case a_t^* is among the bad arms and $\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) > \tilde{\Delta}_\ell(a_t^*)/2$. We show that the change in the mean of a_t^* can be detected relatively quickly.

Let $t \in I_{\ell,i}$. We start by bounding the regret in terms of $\tilde{\Delta}_\ell(a_t^*)$. Let $[s', t']$ be the interval through which a_t^* was evicted from the good arms. Then, by using (3),

$$\begin{aligned}
\mu_t(a_t^*) - \mu_t(a_t^g) &\le \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) + \tilde{\mu}_\ell(a_t^*) - \hat{\mu}_{[\beta_{\ell,i},t]}(a_t^g) + \sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \\
&\le \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) + \hat{\mu}_{[s',t']}(a_t^*) - \hat{\mu}_{[s',t']}(a_t^g) + \sqrt{\frac{2\log T}{n_{[s',t']}(a_t^g)}} + 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \\
&\le \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) + \sqrt{\frac{C_1 \log T}{n_{[s',t']}(a_t^*) - 1}} + 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \\
&\le \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) + \tilde{\Delta}_\ell(a_t^*) + 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} \\
&\le 3[\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)] + 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}}
\end{aligned}$$

for sufficiently large C_1.

For $t \in I_{\ell,i}$, let ϵ_i be the largest $\epsilon = \tilde{\Delta}_\ell(a_t^*)/2^j$ such that $\epsilon \le [\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)]/8$, and $n_i = \lceil (2\log T)/\epsilon_i^2 \rceil$. Then $[\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)] \le 16\epsilon_i$ and the regret is bounded as

$$\mu_t(a_t^*) - \mu_t(a_t^g) \le 48\epsilon_i + 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}}.$$

Since $n_{[\beta_{\ell,i},t]}(a_t^g) \ge \lfloor (t - \beta_{\ell,i} + 1)/K \rfloor$, the overall contribution to the regret of the second term on the right hand side is

$$\sum_{\ell,i} \sum_{t \in I_{\ell,i} \cap \mathcal{B}_{\ell,2}} 2\sqrt{\frac{2\log T}{n_{[\beta_{\ell,i},t]}(a_t^g)}} = \sum_{\ell,i} O\left(\sqrt{K|I_{\ell,i}|\log T}\right) = O\left(\sqrt{KLT\log T}\right).$$

The overall contribution of small ϵ_i, $\epsilon_i \le \sqrt{K(\log T)/|I_{\ell,i}|}$, is also at most $O\left(\sqrt{KLT\log T}\right)$. Thus we consider only $\epsilon_i > \sqrt{K(\log T)/|I_{\ell,i}|}$.

For the first term $48\epsilon_i$, we start with considering intervals of small size, $|I_{\ell,i}| < 2Kn_i$. The contribution of such an interval is bounded as

$$2Kn_i\epsilon_i \le 4K(\log T)/\epsilon_i + K \le 4\sqrt{K|I_{\ell,i}|(\log T)} + K.$$

Summing over all such intervals gives again $O\left(\sqrt{KLT\log T}\right)$.

Finally, we consider the contribution of ϵ_i in large intervals, $|I_{\ell,i}| \ge 2Kn_i$. If the algorithm selects a_t^* in the interval $I_{\ell,i}$ for n_i times, $n_i = n_{[\beta_{\ell,i},t]}(a_t^*)$, then by Assumption 6, the conditions

of $\mathcal{B}_{\ell,2}$, and since $\sqrt{2(\log T)/n_i} \le \epsilon_i \le [\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)]/8$, we have

$$\begin{aligned}\hat{\mu}_{[\beta_{\ell,i},t]}(a_t^*) - \tilde{\mu}_\ell(a_t^*) &> \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) - \sqrt{\frac{2\log T}{n_i}} \\ &\ge \mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*) + \sqrt{\frac{2\log T}{n_i}} - 2\epsilon_i \\ &\ge \frac{3}{4}[\mu_t(a_t^*) - \tilde{\mu}_\ell(a_t^*)] + \sqrt{\frac{2\log T}{n_i}} \ge \tilde{\Delta}_\ell(a_t^*)/4 + \sqrt{\frac{2\log T}{n_i}},\end{aligned}$$

such that condition (4) for starting a new episode is satisfied. Such selections of arm a_t^* are enforced, if the algorithm adds the checking obligation (ϵ_i, n_i, s) to $\mathcal{S}_s(a_t^*)$ for any $\beta_{\ell,i} \le s \le \beta_{\ell,i+1} - Kn_i$.

Let $J_1, \ldots, J_N$, $J_k = [\alpha_k, \alpha_{k+1} - 1]$, be the partition of $[1, T]$ into intervals without change. Then $\mu_{s_1}(a) = \mu_{s_2}(a)$ for $s_1, s_2 \in J_k$ and any arm a, and $\mu_{\alpha_k}(a) \ne \mu_{\alpha_{k+1}}(a)$ for some arm a. Note that each $I_{\ell,i}$ is subset of some J_k, and that the intervals $I_{\ell,i}$ are random and depend on the random rewards observed by the algorithm.

We will prove a bound on the total future expected contributions of the ϵ_i when $t \in \mathcal{B}_{\ell,2}$ and t is in a large interval, starting from the current interval $I_{\ell,i}$ with starting point $\beta_{\ell,i}$. We denote this contribution by $R_{\ell,i}$ which is conditioned on $\beta_{\ell,i}$. Let $I_{\ell,i} \subseteq J_k$. We will show by backward induction that

$$R_{\ell,i} \le \sum_{l=\ell}^{L} \sqrt{\frac{KT\log T}{l}} + 4\sqrt{K(\log T)(2L - k - \ell)(T + 1 - \beta_{\ell,i})}.$$

This is obviously true after all time steps when $\beta_{\ell,i} = T + 1$.

Since we are interested only in large intervals, we assume that $\alpha_{k+1} - \beta_{\ell,i} \ge n_i$. A change is detected if a checking obligation (ϵ_i, n_i, s) is executed, and such a checking obligation is added with probability $p_\ell \epsilon_i$, $p_\ell = \sqrt{\ell/(KT\log T)}$. Thus the contribution $R'_{\ell,i}$ within interval $I_{\ell,i}$ is at most

$$\begin{aligned}R'_{\ell,i} &\le \epsilon_i \left[\sum_{h=1}^{\alpha_{k+1} - \beta_{\ell,i} - Kn_i} (1 - p_\ell\epsilon_i)^h + Kn_i\right] \le \epsilon_i \left[\frac{1 - (1 - p_\ell\epsilon_i)^{\alpha_{k+1} - \beta_{\ell,i} - Kn_i}}{p_\ell\epsilon_i} + Kn_i\right] \\ &= \frac{1}{p_\ell}\left[1 - (1 - p_\ell\epsilon_i)^{\alpha_{k+1} - \beta_{\ell,i} - Kn_i}\right] + Kn_i\epsilon_i \\ &\le \frac{1}{p_\ell}\left[1 - (1 - p_\ell\epsilon_i)^{\alpha_{k+1} - \beta_{\ell,i} - Kn_i}\right] + 4K(\log T)/\epsilon_i.\end{aligned}$$

Let $q_{\ell,i}$ be the probability that episode ℓ does not end within $I_{\ell,i}$, which means that $\beta_{\ell,i+1} = \alpha_{k+1}$. Then

$$q_{\ell,i} \le (1 - p_\ell\epsilon_i)^{\alpha_{k+1} - \beta_{\ell,i} - Kn_i}$$

and by induction

$$\begin{aligned}
R_{\ell,i} &\le R'_{\ell,i} + q_{\ell,i} R_{\ell,i+1} + (1-q_{\ell,i}) R_{\ell+1,1} \\
&\le \frac{1}{p_\ell}\left[1-q_{\ell,i}\right] + 4q_{\ell,i}\sqrt{K(\log T)(\alpha_{k+1}-\beta_{\ell,i})} + 4(1-q_{\ell,i})\sqrt{K(\log T)(\beta_{\ell+1,1}-\beta_{\ell,i})} \\
&\qquad + \frac{q_{\ell,i}}{p_\ell} + \sum_{l=\ell+1}^{L} \frac{1}{p_l} \\
&\qquad + 4q_{\ell,i}\sqrt{K(\log T)(2L-k-\ell-1)(T+1-\alpha_{k+1})} \\
&\qquad + 4(1-q_{\ell,i})\sqrt{K(\log T)(2L-k-\ell-1)(T+1-\beta_{\ell+1,1})} \\
&\le \sum_{l=\ell}^{L} \frac{1}{p_l} + 4\sqrt{K(\log T)(2L-k-\ell)(T+1-\beta_{\ell,i})},
\end{aligned}$$

since $\sqrt{x}+\sqrt{ny} \le \sqrt{(n+1)(x+y)}$. The total contribution of long intervals is bounded by

$$R_{1,1} \le 2\sqrt{KLT\log T} + 4\sqrt{2KLT(\log T)},$$

which concludes the regret analysis.