Journal of Machine Learning Research 24 (2023) 1-37 Submitted 4/22; Revised 9/23; Published 10/23

A New Look at Dynamic Regret for Non-Stationary Stochastic Bandits

Yasin Abbasi-Yadkori YADKORI@GOOGLE.COM
Google DeepMind, London, UK

András György AGYORGY@GOOGLE.COM
Google DeepMind, London, UK

Nevena Lazić NEVENA@GOOGLE.COM
Google DeepMind, Mountain View, USA

Editor: Ambuj Tewari

Abstract

We study the non-stationary stochastic multi-armed bandit problem, where the reward statistics of each arm may change several times during the course of learning. The performance of a learning algorithm is evaluated in terms of its dynamic regret, which is defined as the difference between the expected cumulative reward of an agent choosing the optimal arm in every time step and the cumulative reward of the learning algorithm. One way to measure the hardness of such environments is to consider how many times the identity of the optimal arm can change. We propose a method that achieves, in K-armed bandit problems, a near-optimal $\widetilde{O}(\sqrt{KN(S+1)})$ dynamic regret, where N is the time horizon of the problem and S is the number of times the identity of the optimal arm changes, without prior knowledge of S. Previous works for this problem obtain regret bounds that scale with the number of changes (or the amount of change) in the reward functions, which can be much larger, or assume prior knowledge of S to achieve similar bounds.

Keywords: Online learning, multi-armed bandits, non-stationary learning, dynamic regret, tracking.

1. Introduction

The multi-armed bandit (MAB) problem is the canonical problem for studying the exploration-exploitation dilemma. At each time step $n \in \{1, \dots, N\}$, the learner selects an arm (also called action) $a_n \in \{1, .., K\}$ and receives a reward r_n generated from an unknown distribution which may depend on both the time step and the action. The learner's goal is to maximize the sum of the rewards. In the standard stochastic MAB problem, the reward distribution for each arm is assumed to be stationary, and algorithms are evaluated based on their expected *regret*, which is the difference between the expected rewards obtained by the algorithm and the best fixed arm in hindsight.

In this work, we consider the MAB problem with reward distributions that are non-stationary and can change several times during the course of learning. We evaluate learning algorithms in terms of their dynamic regret, which is the difference between the cumulative expected rewards obtained by the best non-stationary policy selecting the optimal arm in every time step and those of the learning algorithm. MAB problems with non-stationary reward distributions have been studied extensively in literature. This includes a variety of settings, including adversarial rewards, limited total variation of change (Besbes et al., 2014), limited number of switches (Auer et al., 2002), as well as imposing additional assumptions on the process generating the changes (Ortner et al., 2014; Slivkins and Upfal, 2008; Cao et al., 2019).

In general, the achievable dynamic regret depends on the assumptions made about the reward process. When the reward distribution changes at most L times, known as the switching bandit problem (Garivier and Moulines, 2011), the EXP3.S algorithm of Auer et al. (2002) can achieve $O(\sqrt{K(L+1)N\log(KN)})$ regret when its tuning depends on L, which therefore needs to be known in advance.[1] This result is known to be minimax optimal up to the logarithmic factors (Garivier and Moulines, 2011). Several other algorithms can be tuned to achieve bounds that are optimal in N and L given the prior knowledge of L, including the sliding-window UCB algorithm of Garivier and Moulines (2011) and an elimination-based method by Allesiardo et al. (2017).

The first algorithm to obtain near-optimal regret without knowing the number of changes L is the AdSwitch algorithm of Auer et al. (2018). While the original version of AdSwitch is only applicable to the case of $K = 2$, it was subsequently extended to general K by Auer et al. (2019b); Chen et al. (2019); Auer et al. (2019a). The algorithm starts with an initial estimation stage that detects the current optimal arm, and then plays that arm while performing periodic exploration in order to detect changes. Another recent such algorithm is Master (Wei and Luo, 2021), which, specialized to the switching bandit problem, runs a baseline UCB1 algorithm (Auer and Ortner, 2010) at multiple time scales, and periodically resets if non-stationarity is detected.

While the number of times the reward distribution changes is often indeed related to the hardness of the problem, it can be a quite pessimistic measure of complexity: for example, a change in the reward distribution of a suboptimal arm that leaves it suboptimal, or a slight change in the reward of the optimal arm so that it remains optimal should not really affect the performance of good learning algorithms.

To address this issue, in this paper we aim to bound the regret in terms of the number of changes in the identity of the optimal arm S, which can be much smaller than the number of reward changes L considered in prior work (note that EXP3.S of Auer et al., 2002 can readily give a bound which scales with S instead of L, but it requires prior knowledge of S, as discussed above). We propose a modified version of AdSwitch, called ArmSwitch, which performs periodic exploration in order to detect a change in the optimal arm, rather than a change in the reward gap. This allows us to obtain a regret bound which scales as $\widetilde{O}(\sqrt{KN(S+1)})$, without the prior knowledge of S.

Similarly to AdSwitch, our algorithm is based on a phased elimination procedure with restarts. The original phased elimination procedure explores arms uniformly at random until it can be determined that an arm is not optimal, at which point the arm is eliminated and moved to the set of "bad" arms, BAD. The arms in the BAD set are occasionally explored to detect a potential change in the optimal arm. The algorithm restarts the phased elimination when all arms are moved to the BAD set.

[1] The EXP3.S algorithm achieves dynamic regret $O(\sqrt{K(L+1)N\log(KN)})$ with respect to any comparator sequence of arms with L changes. By choosing the comparator sequence to be the sequence of optimal arms, we get the aforementioned bound in our problem.

The phased elimination procedure uses confidence intervals to perform arm eliminations. In our setting, even though rewards can change in an arbitrary fashion, we can construct confidence intervals (based on the idea of importance weighting) for a notion of weighted total rewards. Using such confidence intervals however introduces an issue: it is difficult to control the probability of detecting a change in the optimal arm as the number of active arms can fluctuate in a complicated way, and this can lead to a regret bound that scales worse than the optimal $\sqrt{K}$ scaling in the number of actions K.

In order to resolve this issue, we use a non-uniform sampling distribution in the phased elimination procedure: a fixed probability of $1/K$ is assigned to each active BAD arm[2], while GOOD arms are also selected with equal probability (which is at least $1/K$). This allows us to construct additional confidence intervals with sampling probabilities $1/K$ for each active arm. Using such confidence in addition to ones described above, we can achieve a tighter control on the detection probability, eventually giving the desired overall $\widetilde{O}(\sqrt{KN(S+1)})$ regret bound.

1.1 Notation

For any integer K, $[K] = \{1, \dots, K\}$. For integers $n' < n$, we use $[n' : n]$ to denote the set $\{n', n'+1, \dots, n\}$ (we will often refer to such sets as intervals and use similarly half-open/open intervals not including the corresponding boundary points), and for any sequence $b_{n'}, b_{n'+1}, \dots, b_n$, $b_{n':n} = \sum_{t=n'}^{n} b_t$. We denote the cardinality of a discrete set I by $|I|$. For any two real numbers x, y, $x \vee y = \max\{x, y\}$ denotes their maximum. For an event $\mathcal{E}$, its complement is denoted by $\overline{\mathcal{E}}$, and the indicator $\mathbb{I}\{\mathcal{E}\}$ is 1 if $\mathcal{E}$ holds and zero otherwise.

2. Problem setting

We consider a multi-armed bandit problem with K arms (also called actions) and a known time horizon N (with $K, N \geq 2$). Here, a learning algorithm and an environment interact with each other as follows: At each time step $n \in [N]$, the algorithm selects an arm $A_n \in [K]$ and receives a reward $r_n \in [0, 1]$ drawn according to an unknown distribution $D_n(A_n)$. The goal of the learning algorithm is to collect as much reward as possible.

Let $g_n(a)$ denote the mean of $D_n(a)$ for any $a \in [K]$ and $n \in [N]$. We assume $g_n(a) \in [0, 1]$ and we call it the (mean) reward function. Our assumptions on the rewards and their expectation are made to simplify the presentation and can be replaced with a more general assumption of bounded reward expectation and sub-Gaussian noise. We also assume that the sequence of reward distributions is chosen by an oblivious adversary before the start of the game. Let $a_n^* \in \operatorname{argmax}_{a \in [K]} g_n(a)$ be an optimal arm at time n. We say an arm $a \in [K]$ is optimal in the interval $[n : n']$ if $a \in \operatorname{argmax}_{a \in [K]} g_t(a)$ for all $t \in [n : n']$. We evaluate the performance of algorithms in terms of the dynamic regret, defined with respect to the sequence of optimal arms as

$$R_N := \sum_{n=1}^{N} (g_n(a_n^*) - g_n(A_n)) ,$$

where A_n denotes the arm selected by the algorithm at time n.

[2] By BAD (GOOD) arms, we mean the set of arms that are in the BAD (GOOD) set.

We are interested in bounding the dynamic regret with the number of changes in the identity of the optimal arm. To this end, for any sequence $x_1, \ldots, x_N$ we define

$$S(x_1, \ldots, x_N) := \sum_{n=1}^{N-1} \mathbb{I}\{x_n \neq x_{n+1}\} .$$

and denote by $S = S(a_1^*, \ldots, a_N^*)$ the number of times the identity of the optimal arm changes. Since the optimal arm a_n^* may not be unique for any n, we select $a_1^*, \ldots, a_N^*$ such that they minimize $S(a_1^*, \ldots, a_N^*)$ over the possible sequences of optimal arms (and hence S takes the smallest value possible). Our goal is to design an algorithm that satisfies the following regret bound without the knowledge of S:

$$\mathbb{E}[R_N] = \widetilde{O}\left(\sqrt{KN(S(a_1^*, \ldots, a_N^*) + 1)}\right) . \tag{1}$$

We emphasize that this goal is not addressed in the literature on stochastic non-stationary bandits. In most existing results, the regret bounds depend on the number of changes in the reward distributions D_n instead, which can be much larger than S.

3. Algorithm

To develop our algorithm, we start with the AdSwitch method of Auer et al. (2019b), which was designed for the standard piecewise stationary scenario, where the reward distributions may change only a limited number of times (say L), and they remain constant in between. Conceptually, AdSwitch works the following way. For every stationary segment it maintains a set of arms (called the GOOD set), one of which is with high probability guaranteed to be the optimal arm for the given segment, and these arms are pulled in a round robin fashion. The expected reward of each arm is estimated based on the observations (calculating the average reward for each arm together with confidence intervals), and if the estimates imply with high probability that an arm cannot be optimal, it is removed from the GOOD set. This phased elimination strategy is known to achieve an optimal regret rate for stationary stochastic bandits (Auer and Ortner, 2010). To be able to detect the end of a stationary segment, AdSwitch also explores arms which are not in the GOOD set (the set of those arms is called the BAD set), with a carefully designed exploration strategy, striking a good balance between the exploration probability and the amount of change the selected exploration strategy can detect. If a change is detected (with high probability), the algorithm declares the end of the current segment, and it is restarted, with all the arms being in the GOOD set and all estimates reset.

Our algorithm, called ArmSwitch (for considering the number of times the optimal *arm* changes), is based on similar ideas, but we need to introduce several changes in both the algorithm and its analysis to make it work in our setting. Similarly to AdSwitch, ArmSwitch tries to identify adaptively the segments where the identity of the optimal arm remains the same, and also to learn the optimal arm in each segment. As such, we also maintain a GOOD set of arms, one of which is guaranteed to be the optimal arm for the given segment with high probability, while also carefully exploring arms in the BAD

Algorithm 1 ArmSwitch algorithm

```
Input: time horizon $N$, constant $\delta \in (0,1)$
 1: Initialize $s = 0, n = 0$.
 2: Start a new episode:
    $s \leftarrow s+1$, $t_s \leftarrow n+1$            {** $t_s$ is the start of phase $s$ **}
    GOOD $\leftarrow \{1, \dots K\}$, BAD $\leftarrow \emptyset$.
    $B(a) \leftarrow 0$, Active$(a) \leftarrow n+1$, $\forall a$.      {** $B(a)$ is exploration obligation of arm $a$ **}
 3: Next time step: $n \leftarrow n+1$
    {** History $\mathcal{H}'_n$ is the information available to the algorithm at this point **}
 4: for $a \in$ BAD do
 5:   for $\varepsilon \in \mathcal{B} = \{2^{-1}, 2^{-2}, \dots 2^{-\lceil \log_2 N \rceil}\}$ do
 6:     With probability $\varepsilon/\sqrt{K(n+1-t_s)}$:
 7:     if $B(a) \le 0$ then
 8:       Set Active$(a) \leftarrow n$.
 9:     end if
10:     $B(a) \leftarrow \max(B(a), 1/\varepsilon^2)$.
11:   end for
12: end for
13: Define the active set $\mathcal{A} = \text{GOOD} \cup \{a \in \text{BAD} : B(a) \ge \frac{1}{K}\}$.
14: Set $B(a) \leftarrow 0$ for all $a \notin \mathcal{A}$, and let $m = |\mathcal{A} \cap \text{BAD}|$.
    {** History $\mathcal{H}_n$ is the information available to the algorithm at this point **}
15: Define distribution $P_n$ by $P_n(a) = \frac{1}{K}$ for $a \in \mathcal{A} \cap \text{BAD}$, $P_n(a) = \frac{1-m/K}{|\text{GOOD}|}$ for $a \in$ GOOD,
    and $P_n = 0$ for $a \notin \mathcal{A}$.
16: Select $A_n$ by sampling from $P_n$ and receive reward $r_n$.
    {** Define variables with index $n$: GOOD$_n$ = GOOD, BAD$_n$ = BAD, $B_n(a) = B(a)$,
    etc. **}
17: Set $B(a) \leftarrow B(a) - \frac{1}{K}$ for all $a \in \mathcal{A} \cap \text{BAD}$. {** Exp. obl. consumed in one round **}
18: for $a, a' \in \mathcal{A}$ do
19:   if Elim$_n(a', a)$ then
20:     If $a \in$ GOOD, move $a$ to BAD.
21:     If $a \in$ BAD, set $B(a) \leftarrow 0$.
22:   end if
23:   if GOOD is empty then
24:     Go to 2 (start a new episode)
25:   end if
26: end for
27: Go to 3 (next time step).
```

set.[3] Note, however, that because in our case the reward function can change arbitrarily for every arm, deterministically going over all active arms (the GOOD arms and the ones being explored, together the active arms $\mathcal{A}$) and taking the average observed reward for every

[3] Throughout the paper, when these sets are referred to in a specific time step, they are indexed with that time step (as mentioned in the comment in line 16 of the algorithm), and we also refer to the arms in the GOOD, resp. BAD, set as GOOD, resp. BAD, arms.

Algorithm 2 The $\textsc{Elim}_n(a', a)$ subroutine

```
for $n' \in [\max\{\text{Active}_n(a), \text{Active}_n(a')\}, n]$ do
    if $\widehat{\widehat{\Delta}}_{n':n}(a', a) > 12 C_{n',n} \left( \sqrt{\frac{n-n'+1}{K}} \vee C_{n',n} \right)$ (see definition (3) and condition (6)),
    or $a, a' \in \text{GOOD}$ and $\widehat{\Delta}_{n':n}(a', a) > 12 C_{n',n} \left( \sqrt{P_{n':n}} \vee C_{n',n} \right)$ (see definition (2) and condition (7)) then
        return true
    end if
end for
return false
```

arm would not give good estimates of the arms' performances over the segment. Instead, we randomly sample from the active set $\mathcal{A}$, and when comparing the performance of two arms over an interval where both arms were active, we compare their cumulative weighted reward where each reward is weighted with the sampling probability (instead of the average observed reward) irrespective of how many times the arms were actually used. We provide more details on the sampling and comparison procedures in the next section. Most notably, the sampling distribution is non-uniform and assigns a fixed probability of $1/K$ to any BAD arm in $\mathcal{A}$ (i.e., to arms in $\text{BAD} \cap \mathcal{A}$).

3.1 The ArmSwitch algorithm

Our algorithm is shown in Algorithm 1. It is an elimination algorithm with repeated exploration and restarts. The algorithm proceeds in episodes $s = 1, 2, ...$; we use t_s to denote the start time of the s^{th} episode.

The algorithm continuously maintains a set of arms GOOD containing the arms which have not been ruled out to be optimal in the current episode by some statistical test, while $\text{BAD} = [K] \setminus \text{GOOD}$ contains all other arms. In what follows, we use GOOD_n and BAD_n to denote the GOOD and BAD sets at time n, respectively. At the beginning of each episode, GOOD contains all the arms, which are then eliminated (and moved to BAD) when it can be proved with high probability based on the received rewards that they cannot be optimal for at least one time step of the episode. When all arms are eliminated from the GOOD set, the algorithm knows that the identity of the optimal arm has changed with high probability, so a new episode is started (where again all arms can be optimal initially). An algorithm that only plays the arms in the GOOD set would miss detecting if an arm from the BAD set became good. To handle such cases, in every step ArmSwitch may select, with some small probability, some arms from the BAD set to be explored, and the algorithm may play these arms as well, beside the ones in the GOOD set; the arms which can be played in any given time step are collected in the active set $\mathcal{A}$.

Next we describe each component of ArmSwitch in detail:

Exploration of arms in the BAD set. To facilitate exploration of arms in the BAD set, ArmSwitch maintains so-called exploration obligations $B(a)$, containing a prescribed sum of probabilities with which a bad arm has to be explored over multiple time steps (this is similar to exploration obligations used by AdSwitch prescribing how many times an arm has to be sampled, but it is better tuned for random sampling). When any arm gets to the BAD set, its exploration obligation is set to 0 (lines 20 and 21). At the beginning of a time step n (before an arm is played, see lines 4–12), the algorithm may schedule some exploration for some arms in the BAD set: The obligation (length) can be $1/\varepsilon^2$ for any ε in an exponential grid $\mathcal{B} = \{1/2, 1/4, \ldots, 2^{-\lceil \log_2 N \rceil}\}$, and longer explorations have smaller probabilities: in episode s, $a \in \mathrm{BAD}_n$ and $\varepsilon \in \mathcal{B}$, with probability $\varepsilon/\sqrt{K(n+1-t_s)}$, we prescribe an exploration obligation of length $1/\varepsilon^2$ by setting $B(a)$ to $\max(B(a), 1/\varepsilon^2)$. We say that an obligation is *scheduled* in a time step for arm a if it is the longest obligation prescribed, and it is larger than the previous obligation, and define a corresponding event $\mathrm{EXP}(a, n, \varepsilon)$ which holds if and only if an exploration obligation of length $1/\varepsilon^2$ is scheduled for arm a in time step n (i.e., $B_n(a) = 1/\varepsilon^2 > B_{n-1}(a)$). Note that the conditional probability of $\mathrm{EXP}(a, n, \varepsilon)$ (given the history up to this point) if n belongs to episode s is at most $\varepsilon/\sqrt{K(n+1-t_s)}$.

Arms in the BAD set with positive exploration obligations typically belong to the active set $\mathcal{A}$ for the given time step, unless their exploration obligation is "rounded down" to 0 (see lines 13–14, and the description of the procedure for sampling an arm below). After the algorithm plays an arm and receives a reward, the exploration obligations for any active arm in the BAD set are reduced by the (conditional) probability of selecting that arm, that is, by $\frac{1}{K}$ (line 17).

Selecting arms. After possibly introducing new sampling obligations, the algorithm selects a set of active arms $\mathcal{A}$ from which the played arm A_n is selected. Set $\mathcal{A}$ contains all the arms in the GOOD set, and also all the BAD arms with exploration obligations at least $\frac{1}{K}$. We also use $\mathcal{A}_n$ to denote the active arms at time n. The algorithm selects an action from a distribution that assigns probability $\frac{1}{K}$ to any BAD arm in the active set, and is uniform over the GOOD arms using the remaining probabilities (line 15). It will be helpful to define an additional variable $\widetilde{A}_n$ by implementing the sampling of A_n in a two-step procedure: with probability $\frac{|\mathcal{A}|}{K}$, an arm is sampled uniformly at random from $\mathcal{A}$, and otherwise an arm is sampled uniformly at random from GOOD. The sampled arm is denoted by A_n; if the first event happens, we also let $\widetilde{A}_n := A_n$, while in case of the second event, we let $\widetilde{A}_n$ to take a value not in the action set by defining $\widetilde{A}_n := *$ (an alternative definition of $\widetilde{A}_n$ is to choose A_n according to P_n and then set $\widetilde{A}_n$ to A_n with probability $\frac{1}{K}/P_n(A_n)$ and to $*$ otherwise).

Eliminating arms from the active set. If, based on the observed rewards r_n and the selected actions, the algorithm can prove (with high probability) that an arm in the active set $\mathcal{A}$ cannot be optimal starting from the last time it has become active, it is removed from the active set, that is, it is removed from the set of GOOD arms if it belonged there, or its exploration obligations are deleted if it belongs to the BAD set (lines 18–22). This elimination is based on observations in intervals when an arm is *active*: we say that arm a is active in time step n if $a \in \mathcal{A}_n$. The start of the most recent active period is maintained in the variable Active: for arms in the GOOD set, it is the beginning of the current episode (line 2), while for active arms a in the BAD set, it is the time when the algorithm has decided to explore them (i.e., when the exploration obligation $B(a)$ last became positive, see line 8). If $a \in \mathcal{A}_n$, then a is active in $[n' : n]$ if and only if $\mathrm{Active}_n(a) \le n'$. We denote by the boolean $\textsc{Elim}_n(a', a)$ whether we can prove (with high probability) in time step n

that arm a' was better in at least one time step than a during a time interval ending at time n when both arms were active. In this case we know that a cannot be an optimal arm in this interval, hence it cannot be an optimal arm in the episode, and so it can be eliminated. The details of this procedure, which is at the heart of ArmSwitch, are given in the next section (Section 3.2) and Algorithm 2.

3.2 Comparison of arms ($\mathbf{Elim}_n(a', a)$)

We now specify the arm-elimination condition $\textsc{Elim}_n(a', a)$, which indicates whether arm a' is better than a in at least one time step in the current episode up to time n, with high probability, given the algorithm's history up to that point. If it is the case that a' is better than a, arm a can be eliminated from the active set by a'. The formal definition is given in Algorithm 2. The notation used in the algorithm and its analysis is introduced below and summarized in Table 1.

Let $P_n(a)$ denote the probability of selecting arm a in time step n, given the history $\mathcal{H}_n$ up to that point (see line 14 of Algorithm 1). By definition, $P_n(a) = \frac{1}{K}$ for $a \in \mathcal{A}_n \cap \mathrm{BAD}_n$, and $P_n(a) = \frac{1}{|\mathrm{GOOD}_n|}\left(1 - \frac{|\mathcal{A}_n \cap \mathrm{BAD}_n|}{K}\right)$ for $a \in \mathrm{GOOD}_n$ (note that the latter is always at least $1/K$). For $a \notin \mathcal{A}_n$, $P_n(a) = 0$. Let A_n be the arm selected at time n, $G_n(a) := P_n(a) g_n(a)$ be the weighted expected reward of arm a in time step n, and $\widehat{G}_n(a) := \mathbb{I}\{A_n = a\} r_n$ be its estimate; note that this is an unbiased estimate since $G_n(a) = \mathbb{E}[\widehat{G}_n(a) | \mathcal{H}_n]$. We also define $\widetilde{G}_n(a) := \frac{g_n(a)}{K}$ and its estimate $\widehat{\widetilde{G}}_n(a) := \mathbb{I}\{\widetilde{A}_n = a\} r_n$ (recall that $\widetilde{A}_n$ equals A_n with probability $|\mathcal{A}_n|/K$ and $*$ otherwise, see its definition in the paragraph on selecting arms in Section 3.1). For any interval $[n' : n]$, we consider the following weighted sums of (possibly expected) rewards:

$$G_{n':n}(a) := \sum_{t \in [n':n]} P_t(a) g_t(a) , \qquad \widehat{G}_{n':n}(a) := \sum_{t \in [n':n]} \widehat{G}_t(a) = \sum_{t \in [n':n]} \mathbb{I}\{A_t = a\} r_t ,$$

$$\widetilde{G}_{n':n}(a) := \frac{1}{K} \sum_{t \in [n':n]} g_t(a) , \qquad \widehat{\widetilde{G}}_{n':n}(a) := \sum_{t \in [n':n]} \widehat{\widetilde{G}}_t(a) = \sum_{t \in [n':n]} \mathbb{I}\{\widetilde{A}_t = a\} r_t .$$

Note that $\widetilde{G}_{n':n}(a) = G_{n':n}(a)$ and $\widehat{\widetilde{G}}_{n':n}(a) = \widehat{G}_{n':n}(a)$ for $a \in \bigcap_{t=n'}^{n} \mathcal{A}_t \cap \mathrm{BAD}_t$.[4] Define $\Delta_{n':n}(a, a')$, $\widehat{\Delta}_{n':n}(a, a')$, $\widetilde{\Delta}_{n':n}(a, a')$ and $\widehat{\widetilde{\Delta}}_{n':n}(a, a')$ as follows: for any arms $a, a' \in [K]$, let

$$\Delta_{n':n}(a, a') := G_{n':n}(a) - G_{n':n}(a') , \qquad \widehat{\Delta}_{n':n}(a, a') := \widehat{G}_{n':n}(a) - \widehat{G}_{n':n}(a') , \tag{2}$$

$$\widetilde{\Delta}_{n':n}(a, a') := \widetilde{G}_{n':n}(a) - \widetilde{G}_{n':n}(a') , \qquad \widehat{\widetilde{\Delta}}_{n':n}(a, a') := \widehat{\widetilde{G}}_{n':n}(a) - \widehat{\widetilde{G}}_{n':n}(a') . \tag{3}$$

Note that $\widetilde{\Delta}_{n',n}(a, a') = -\widetilde{\Delta}_{n',n}(a', a)$ and $\widehat{\widetilde{\Delta}}_{n',n}(a, a') = -\widehat{\widetilde{\Delta}}_{n',n}(a', a)$. Since for an active arm a, $\widehat{\widetilde{G}}_n(a)$ is an unbiased estimate of $\widetilde{G}_n(a)$ as explained above, using martingale concentration it can be shown that $\widehat{\widetilde{G}}_{n',n}(a)$ is close to $\widetilde{G}_{n',n}(a)$ (and in turn $\widehat{\widetilde{\Delta}}_{n',n}(a, a')$ is

[4] Also notice that since $\mathbb{I}\{A_t = a\}$ is a Bernoulli random variable whose variance is upper bounded by $P_t(a)$, $G_{n':n}(a)$ is similar in spirit to the minimum-variance estimator of a joint mean of independent random variables, where the optimal weighting is proportional to the variance of each variable.

Probabilities			
$P_{n':n}(a)$	$\sum_{t=n'}^{n} P_t(a)$		
Confidence intervals			
$C_{n',n}$	$\sqrt{\log\left(\frac{2KN^2(\log(n-n'+1)+2)}{\delta}\right)}$	$\boxed{C_N = C_{1,N}}$	
$C'_{n',n}$	$\sqrt{\log\left(\frac{2K^2N^2(\log(n-n'+1)+2)}{\delta}\right)}$		
Rewards		Gaps	
$G_n(a)$	$P_n(a)g_n(a)$	$\Delta_{n':n}(a,a')$	$G_{n':n}(a) - G_{n':n}(a')$
$\widehat{G}_n(a)$	$\mathbb{I}\{A_n = a\}r_n$	$\widehat{\Delta}_{n':n}(a,a')$	$\widehat{G}_{n':n}(a) - \widehat{G}_{n':n}(a')$
$\widetilde{G}_n(a)$	$\frac{g_n(a)}{K}$	$\widetilde{\Delta}_{n':n}(a,a')$	$\widetilde{G}_{n':n}(a) - \widetilde{G}_{n':n}(a')$
$\widehat{\widetilde{G}}_n(a)$	$\mathbb{I}\{\widetilde{A}_n = a\}r_n$	$\widehat{\widetilde{\Delta}}_{n':n}(a,a')$	$\widehat{\widetilde{G}}_{n':n}(a) - \widehat{\widetilde{G}}_{n':n}(a')$
$G_{n':n}(a)$	$\sum_{t\in[n':n]} G_t(a)$	$\widetilde{\Delta'}_{n':n}(a,a')$	$\sum_{t=n'}^{n} \mathbb{I}\{\widetilde{A}_t = a\}\big(g_t(a) - g_t(a')\big)$
$\widehat{G}_{n':n}(a)$	$\sum_{t\in[n':n]} \widehat{G}_t(a)$		
$\widetilde{G}_{n':n}(a)$	$\sum_{t\in[n':n]} \widetilde{G}_t(a)$		
$\widehat{\widetilde{G}}_{n':n}(a)$	$\sum_{t\in[n':n]} \widehat{\widetilde{G}}_t(a)$		
$G_{n':n}(a,a')$	$\sum_{t=n'}^{n} P_t(a)g_t(a')$		
$G'_{n':n}(a,a')$	$\sum_{t=n'}^{n} \mathbb{I}\{A_t = a\}g_t(a')$		
Events			
$\mathcal{E}_1$	$\left\lvert\widehat{G}_{n':n}(a) - G_{n':n}(a)\right\rvert \le 6C_{n',n}\left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right)$ for all $[n':n] \subseteq [N], a \in [K]$		
$\mathcal{E}_2$	$\left\lvert\widehat{\widetilde{G}}_{n':n}(a) - \widetilde{G}_{n':n}(a)\right\rvert \le 6C_{n',n}\left(\sqrt{\frac{n-n'+1}{K}} \vee C_{n',n}\right)$ for all $[n':n] \subseteq [N]$ and $a \in [K]$ active on $[n':n]$		
$\mathcal{E}_3$	$\left\lvert G'_{n':n}(a,a') - G_{n':n}(a,a')\right\rvert \le 5C'_{n',n}\left(\sqrt{P_{n':n}(a)} \vee C'_{n',n}\right)$ for all $[n':n] \subseteq [N]$ and $a, a' \in [K]$		
$\mathcal{E}_4$	$\left\lvert\widetilde{\Delta'}_{n':n}(a,a') - \widetilde{\Delta}_{n':n}(a,a')\right\rvert \le 5C'_{n',n}\left(\sqrt{\frac{n-n'+1}{K}} \vee C'_{n',n}\right)$ for all $[n':n] \subseteq [N]$ and $a, a' \in [K]$ active on $[n':n]$		

Table 1: Summary of notation for the ArmSwitch algorithm and its analysis.

close to $\widetilde{\Delta}_{n',n}(a,a')$) for active arms: defining $C_{n',n} = \sqrt{\log\left(\frac{2KN^2(\log(n-n'+1)+2)}{\delta}\right)}$ for some $\delta \in (0,1)$ and $P_{n':n}(a) = \sum_{t=n'}^{n} P_t(a)$, Lemma 3 shows that

$$\left|\widehat{G}_{n':n}(a) - G_{n':n}(a)\right| \le 6\, C_{n',n}\left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right) , \tag{4}$$

$$\left|\widehat{\widetilde{G}}_{n':n}(a) - \widetilde{G}_{n':n}(a)\right| \le 6\, C_{n',n}\left(\sqrt{\frac{n-n'+1}{K}} \vee C_{n',n}\right) \tag{5}$$

with probability at least $1-\delta$ simultaneously for all intervals $[n':n] \subset [N]$ and actions a that are active on that interval.

If a is an optimal arm in the interval $[n':n]$, then $g_t(a) \ge g_t(a')$ for any other arm a' and $t \in [n':n]$. Therefore, if both a and a' are active in $[n':n]$, $\widetilde{\Delta}_{n',n}(a,a') \ge 0$. Thus, if for an arm a there exists another arm a' such that $\widetilde{\Delta}_{n',n}(a,a') < 0$ — or equivalently, $\widetilde{\Delta}_{n',n}(a',a) > 0$ — then a cannot be optimal, and hence can be eliminated from the set of potentially optimal arms for $[n':n]$. We use our empirical estimates $\widehat{\widetilde{\Delta}}_{n':n}(a,a')$ to verify, with high probability, if this happens: if (5) holds for a, a', then

$$\widehat{\widetilde{\Delta}}_{n':n}(a',a) > 12 C_{n',n}\left(\sqrt{\frac{n-n'+1}{K}} \vee C_{n',n}\right) \tag{6}$$

implies that $\widetilde{\Delta}_{n':n}(a',a) > 0$. The indicator $\textsc{Elim}_n(a',a)$ is true (see Algorithm 2) if (6) holds for any interval $[n':n] \subset [t_s:n]$ such that both a and a' are active on $[n':n]$. The above condition discards data from any time step t such that $\widetilde{A}_t \ne A_t$. Although this way of using data looks sub-optimal, condition (6) does not involve P_t variables, which proves crucial in showing tight regret bounds when the optimal arm is in the BAD set. We also present a variant of our algorithm where the constant 12 in the elimination condition (6) is replaced with 13, implying that $K\widetilde{\Delta}_{n':n}(a',a) > \sqrt{K(n-n'+1)}$.

With a similar argument, we can show that if (4) holds for $a, a' \in \mathrm{GOOD}_n$, then

$$\widehat{\Delta}_{n':n}(a',a) > 12 C_{n',n}\left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right) \tag{7}$$

implies that $\Delta_{n':n}(a',a) > 0$. The indicator $\textsc{Elim}_n(a',a)$ is true if (7) holds for any interval $[n':n] \subset [t_s:n]$ such that both a and a' are in the GOOD set. Further intuition on why we use two elimination conditions are given in Section 4.2.

3.3 Regret of ArmSwitch

The following theorem shows that the regret of ArmSwitch behaves as desired.

Theorem 1 *For a switching multi-armed bandit problem with $K \ge 2$ arms, horizon $N \ge 2$, and $S \ge 0$ changes in the identity of the optimal arm, the expected regret of* ArmSwitch *run with $\delta = 1/(\sqrt{K}N^{3/2})$ is bounded as*

$$\mathbb{E}[R_N] \le \textsc{Const}\sqrt{K(S+1)N}(\log(KN\log(N)))^{3/2}$$

for an appropriate universal constant Const.

If the elimination condition (6) is changed a little bit as indicated above, we can also prove a bound in terms of how much the rewards of the arms can change from one round to the next, usually referred to in the literature as a *variational bound* (Besbes et al., 2014). To this end, let

$$V := \sum_{n=1}^{N-1} \max_{a \in [K]} |g_{n+1}(a) - g_n(a)|$$

be the *total variation* of the reward vectors. The next result is an easy consequence of some intermediate bound obtained in the proof of Theorem 1:

Corollary 2 *Assume* ArmSwitch *is run with condition*

$$\widehat{\widehat{\Delta}}_{n':n}(a', a) > 13 C_{n',n} \left(\sqrt{\frac{n - n' + 1}{K}} \vee C_{n',n} \right) \tag{8}$$

instead of (6) *in line 2 of* $\textsc{Elim}_n(a', a)$ *(Algorithm 2). Then, for an appropriate universal constant* Const *independent of* K, N, V*, the regret of* ArmSwitch *run with* $\delta = 1/(\sqrt{K} N^{3/2})$ *is bounded as*

$$\mathbb{E}[R_N] \leq \textsc{Const} \left(\sqrt{KN} + (KV)^{1/3} N^{2/3} \right) \left(\log \left(KN \log(N) \right) \right)^{3/2}.$$

The rest of the paper (Section 4) is devoted to the proof of the above results.

Remark. In a parallel work, Suk and Kpotufe (2022) proposed a similar algorithm and regret bounds. Compared to ours, the main difference is that their algorithm uses a fully synchronized exploration of all arms, while we also randomize which arms to explore (i.e., when they explore, they explore all BAD arms, that is, completely restart the algorithm in an exploration phase, while we can explore BAD arms individually). This additional randomization introduces some complications in the proof as we need to keep track of the exploration status of each arm separately. Otherwise the resulting algorithms and proof techniques are essentially the same. The elimination condition of Suk and Kpotufe (2022) is slightly different as is designed to detect *significant changes* in the reward distribution (discussed in more details in Section 4.5) instead of changes in the identity of optimal arms; hence, as a parallel to Theorem 1, they prove a bound which scales with the number of significant shifts and not the number of changes to the identity of the optimal arm. They also show a variational bound as given in Corollary 2 above; in fact, we use their definition of significant reward changes to establish our variational bound. More details are given in Section 4.5, along with the proof of the corollary.

4. Analysis

We start with a few useful lemmas then analyze the regret in Section 4.2.

4.1 Useful lemmas

Fix $\delta \in (0,1)$. We define events $\mathcal{E}_1$, $\mathcal{E}_2$ under which the estimates $\widehat{G}_{n',n}(a)$, $\widehat{\widetilde{G}}_{n',n}(a)$ are good:

$$\mathcal{E}_1 = \Bigg\{ \left|\widehat{G}_{n':n}(a) - G_{n':n}(a)\right| \le 6\, C_{n',n} \left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right) \\ \text{for all } [n':n] \subseteq [N] \text{ and actions } a \in [K] \Bigg\} ,$$

$$\mathcal{E}_2 = \Bigg\{ \left|\widehat{\widetilde{G}}_{n':n}(a) - \widetilde{G}_{n':n}(a)\right| \le 6\, C_{n',n} \left(\sqrt{\frac{n-n'+1}{K}} \vee C_{n',n}\right) \\ \text{for all } [n':n] \subseteq [N] \text{ and actions } a \text{ active on } [n':n] \Bigg\} .$$

To help the analysis, it will also be useful to consider the following quantities for any $n' \le n$ and arms $a, a' \in [K]$:

$$G_{n':n}(a,a') := \sum_{t=n'}^{n} P_t(a) g_t(a') \ , \qquad G'_{n':n}(a,a') := \sum_{t=n'}^{n} \mathbb{I}\{A_t = a\} g_t(a') \ ,$$

$$\widetilde{\Delta'}_{n':n}(a,a') := \sum_{t=n'}^{n} \mathbb{I}\{\widetilde{A}_t = a\}\big(g_t(a) - g_t(a')\big) \ .$$

By taking conditional expectations it follows that $G'_{n':n}(a,a')$ should be close to $G_{n':n}(a,a')$ for all arms a, a', and $\widetilde{\Delta'}_{n':n}(a,a')$ should be close to $\widetilde{\Delta}_{n':n}(a,a')$ for arms a, a' that are active on interval $[n':n]$. Defining $C'_{n',n} = \sqrt{\log\left(\frac{2K^2N^2(\log(n-n'+1)+2)}{\delta}\right)}$ (which satisfies $C'_{n',n} \le 2C_{n',n}$), these are formalized in the following events:

$$\mathcal{E}_3 = \Bigg\{ \left|G'_{n':n}(a,a') - G_{n':n}(a,a')\right| \le 5\, C'_{n',n} \left(\sqrt{P_{n':n}(a)} \vee C'_{n',n}\right) \\ \text{for all } [n':n] \subseteq [N] \text{ and actions } a, a' \in [K] \Bigg\} ,$$

$$\mathcal{E}_4 = \Bigg\{ \left|\widetilde{\Delta'}_{n':n}(a,a') - \widetilde{\Delta}_{n':n}(a,a')\right| \le 5\, C'_{n',n} \left(\sqrt{\frac{n-n'+1}{K}} \vee C'_{n',n}\right) \\ \text{for all } [n':n] \subseteq [N] \text{ and actions } a, a' \text{ active on } [n':n] \Bigg\} .$$

The next lemma shows that $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$, and $\mathcal{E}_4$ hold with high probability. Its proof, presented in Appendix A, is based on a version of Freedman's inequality.

Lemma 3 *Each of the events $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$, and $\mathcal{E}_4$ hold with probability at least $1-\delta$.*

We bound the regret of ArmSwitch under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$. Taking into account that the regret in every time step is at most 1 and choosing $\delta = 1/(\sqrt{K}N^{3/2})$, the contribution to the expected regret when $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ or $\mathcal{E}_4$ do not hold can be bounded by $4/\sqrt{KN}$. For arms a, a', note that since

$$\widehat{\Delta}_{n':n}(a', a) = \Big(\widehat{G}_{n':n}(a') - G_{n':n}(a')\Big) + \Big(G_{n':n}(a') - G_{n':n}(a)\Big) + \Big(G_{n':n}(a) - \widehat{G}_{n':n}(a)\Big) ,$$

$\mathcal{E}_1$ implies

$$\left|\widehat{\Delta}_{n':n}(a', a) - \Delta_{n':n}(a', a)\right| \le 6\, C_{n',n} \left(\sqrt{P_{n':n}(a)} \vee C_{n',n} + \sqrt{P_{n':n}(a')} \vee C_{n',n}\right) .$$

In particular, if both a and a' are GOOD throughout the interval $[n' : n]$, then $P_t(a) = P_t(a')$ for all $t \in [n' : n]$, and hence

$$\left|\widehat{\Delta}_{n':n}(a', a) - \Delta_{n':n}(a', a)\right| \le 12\, C_{n',n} \left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right) . \tag{9}$$

Similarly, $\mathcal{E}_2$ implies, for any $a, a' \in [K]$ that are active on $[n' : n]$,

$$\left|\widehat{\widetilde{\Delta}}_{n':n}(a', a) - \widetilde{\Delta}_{n':n}(a', a)\right| \le 12\, C_{n',n} \left(\sqrt{\frac{n-n'+1}{K}} \vee C_{n',n}\right) . \tag{10}$$

The next lemma, proved in Appendix A, shows some useful connections between events $\mathcal{E}_1, \mathcal{E}_2$, the gaps, and $\textsc{Elim}_n(a', a)$.

Lemma 4 *Let $n' \le n$ be two time steps belonging to the same episode.*

(i) Suppose $\mathcal{E}_1$ holds. If $\Delta_{n':n}(a', a) > 24\, C_{n',n}\left(\sqrt{P_{n':n}} \vee C_{n',n}\right)$ for $a, a' \in GOOD_n$, then $\textsc{Elim}_n(a', a)$ is true. Furthermore, if $a, a' \in GOOD_{n+1}$, then $\textsc{Elim}_n(a', a)$ is false and $\Delta_{n':n}(a', a) \le 24\, C_{n',n}\left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right)$.

(ii) Assume $\mathcal{E}_2$ holds and arms a, a' are active on interval $[n' : n]$. If $\widetilde{\Delta}_{n':n}(a', a) > 24\, C_{n',n}\left(\sqrt{\frac{n-n'+1}{K}} \vee C_{n',n}\right)$, then $\textsc{Elim}_n(a', a)$ is true. Furthermore, if $a' \in GOOD_{n+1}$ and either $a \in GOOD_{n+1}$ or $a \in BAD_{n+1} \cap \mathcal{A}_{n+1}$ with no new exploration obligation scheduled for a in time step $n+1$, then $\textsc{Elim}_n(a', a)$ is false and $\widetilde{\Delta}_{n':n}(a', a) \le 24\, C_{n',n}\left(\sqrt{\frac{n-n'+1}{K}} \vee C_{n',n}\right)$.

4.2 Preliminaries to the proof of Theorem 1

We bound the regret of ArmSwitch under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$. We first show that the number of episodes produced by our algorithm is at most $S+1$.

Lemma 5 *If $\mathcal{E}_1 \cap \mathcal{E}_2$ holds,* ArmSwitch *has at most $S+1$ episodes.*

Proof At the beginning of each episode s, the set GOOD_{t_s} includes all arms including the current optimal arm. Under the event $\mathcal{E}_1 \cap \mathcal{E}_2$ and given the elimination conditions (6) and (7), the arm that is optimal at the start of the episode cannot be eliminated from the GOOD set for as long as it stays optimal. Thus, the GOOD set can only become empty after a change in the optimal arm, which happens S times. ■

In what follows, let a^g_s denote an arm that stays in the GOOD set during the entire episode s (i.e., the last arm to get eliminated, or one of the last arms if multiple arms are eliminated when GOOD becomes empty). Let $\mathcal{N}_s$ denote the set of time steps in episode s. We decompose the regret in each episode s into the regret of played arms with respect to a^g_s, and the regret of a^g_s with respect to the optimal arms as follows:

$$R_N = \sum_{n=1}^{N} (g_n(a^*_n) - g_n(A_n)) = R^{(1)}_N + R^{(2)}_N + R^{(3)}_N ,$$

where we define

$$R^{(1)}_N := \sum_{s=1}^{S+1} \sum_{n \in \mathcal{N}_s} \mathbb{I}\{A_n \in \text{BAD}_n\}(g_n(a^g_s) - g_n(A_n)) ,$$

$$R^{(2)}_N := \sum_{s=1}^{S+1} \sum_{n \in \mathcal{N}_s} \mathbb{I}\{A_n \in \text{GOOD}_n\}(g_n(a^g_s) - g_n(A_n)) ,$$

$$R^{(3)}_N := \sum_{s=1}^{S+1} \sum_{n \in \mathcal{N}_s} (g_n(a^*_n) - g_n(a^g_s)) .$$

We proceed by bounding each of the above regret terms. Bounding $R^{(1)}_N$ and $R^{(2)}_N$ can be done using arguments similar to existing techniques of Auer et al. (2019b); this is presented in Section 4.3, in Lemma 6 and Lemma 7, respectively.

The main challenge is controlling $R^{(3)}_N$, and ensuring that it scales properly with the number of arms K is the source of most complications in the algorithm design: using a non-uniform sampling distribution over the active arms, and having two elimination conditions (namely, equations 6 and 7). Indeed, choosing the sampling distribution to be uniform over the active set, which is similar to the round robin action selection of Auer et al. (2019b), would lead to a suboptimal bound in K according to our analysis, as discussed below: Assume that the optimal arm is fixed and active on interval $[n' : n]$, and it belongs to the BAD set. Then condition (7) for eliminating a^g_s (and starting a new episode) is of the form (neglecting the lower order term $C^2_{n',n}$)

$$\sum_{t=n'}^{n} \frac{1}{|\mathcal{A}_t|}(g_t(a^*_t) - g_t(a^g_s)) \approx \sum_{t=n'}^{n} (\widehat{G}_t(a^*_t) - \widehat{G}_t(a^g_s)) \gtrapprox 12\, C_N \sqrt{\sum_{t=n'}^{n} \frac{1}{|\mathcal{A}_t|}} , \qquad (11)$$

where $C_N := C_{1,N}$ which equals $\sqrt{\log(2K^{3/2}N^{7/2}(\log(N)+2))} = \tilde{O}(1)$ for $\delta = 1/(\sqrt{K}N^{3/2})$. To detect by time n that a^*_n is better than a^g_n, we need to explore it for sufficiently long, and with sufficiently high probability. However, $|\mathcal{A}_t|$ can be as small as 2 and as large as K, and so for the same reward sequence, a_n^* needs to be explored $\Theta(\sqrt{K})$ times longer to detect a change if $|\mathcal{A}_t|$ is close to K than if it is close to 2, also resulting in a K times larger regret (assuming the difference between $g_t(a_n)$ and $g_t(a_s^g)$ is constant in $[n', n]$). Therefore, there is an ambiguity (of a factor K) here in selecting the length of the exploration interval, or equivalently, selecting the probability of adding different exploration obligations.

On the other hand, using an action selection probability $1/K$ for actions in BAD, the elimination condition becomes (6), or in simplified form

$$\sum_{t=n'}^{n} (g_t(a_t^*) - g_t(a_s^g)) \gtrapprox 12\, C_N \sqrt{K(n-n'+1)}\ .$$

This in turn allows to set the probability of introducing exploration obligations of different lengths properly, leading to an optimal (up to logarithmic factors) $\widetilde{O}(\sqrt{KSN})$ regret bound. This is proved formally in Lemma 8 in Section 4.4.

Combining Lemmas 6–8 trivially yields Theorem 1. Following the proofs of these lemmas, Corollary 2 is proved in Section 4.5.

4.3 Bounding $R_N^{(1)}$ and $R_N^{(2)}$

We start with bounding $\mathbb{E}[R_N^{(1)}]$.

Lemma 6 *Under the conditions of Theorem 1, we have*

$$\begin{aligned}\mathbb{E}[R_N^{(1)}] &\le 160\, C_N(\log_2(N)+1)\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_s \sqrt{K(t_{s+1}-t_s)}\right] \\ &\quad + \frac{320 C_N(\log_2(N)+1)}{\sqrt{N}} + \frac{2}{\sqrt{KN}} \\ &\le 160\, C_N(\log_2(N)+1)\sqrt{K(S+1)N} + \frac{320 C_N(\log_2(N)+1)}{\sqrt{N}} + \frac{2}{\sqrt{KN}}\end{aligned}$$

where $C_N := C_{1,N}$.

Proof Here we bound the regret due to exploring arms in the BAD set in each episode with respect to the arm a_s^g. Note that all sampling obligations start and end in the same episode, since all arms are placed in the GOOD set at the beginning of each episode. Since the bound trivially holds if $K > N^2$ or $N < 64$, in the rest of the proof we assume $K \le N^2$ and $N \ge 64$.

Consider a sampling obligation for arm $a \in \mathrm{BAD}_n$ that is scheduled in time step n in episode s with exploration parameter ε. Let $\tau_n(\varepsilon, a)$ be the time at which this particular sampling obligation expires, that is, either $n' = \tau_n(\varepsilon, a) + 1$ is the first time after n when the obligation $B_{n'}(a)$ is zero (i.e., it is zero after line 14 of Algorithm 1), or $n' = \tau_n(\varepsilon, a) + 1$ is the first time step when a new exploration obligation is scheduled for a (by triggering the max operation in line 10). Then, under $\mathcal{E}_2 \cap \mathcal{E}_4$, the regret with respect to a_s^g can be bounded as

$$\sum_{t \in [n:\tau_n(\varepsilon,a)]} \mathbb{I}\{A_t = a\}(g_t(a_s^g) - g_t(a))$$

$$\le \widetilde{\Delta}_{n:\tau_n(\varepsilon,a)}(a_s^g, a) + 10C_N \left(\sqrt{\frac{\tau_n(\varepsilon,a) - n + 1}{K}} \vee 2C_N \right) \tag{12}$$

$$\le 34C_N \left(\sqrt{\frac{\tau_n(\varepsilon,a) - n + 1}{K}} \vee C_N \right) + 20C_N^2 + 1$$

$$\le 34C_N \left(\frac{1}{\varepsilon} \vee C_N \right) + 20C_N^2 + 1\,, \tag{13}$$

where the first inequality holds by the definition of $\mathcal{E}_4$ because a is a BAD arm in $[n : \tau_n(\varepsilon, a)]$ (hence $A_t = a$ is equivalent to $\widetilde{A}_t = a$) and $C'_{n',n} \le 2C_N$; the second inequality holds by Lemma 4(ii) (since we assumed that $\mathcal{E}_2$ holds) and the fact that the reward difference between any two arms in one step is at most 1; and the last inequality holds by the fact that the length of the interval cannot be more than K/ε^2 (see line 17 of Algorithm 1).

Recall that $\mathcal{N}_s$ denotes the set of time steps in episode s and that $\mathrm{EXP}(a, n, \varepsilon)$ denotes the event that an exploration obligation of length $1/\varepsilon^2$ is scheduled for arm a in time step n. Note that the exploration intervals $[n, \tau_n(\varepsilon_n, a)] \subset \mathcal{N}_s$ for the scheduled explorations of arm a (that is, for which $\mathrm{EXP}(a, n, \varepsilon_n)$ hold) are disjoint for all n, and together cover all time steps in episode s where a belongs to the BAD set and is active.

Using these observations, the expected regret due to exploring BAD arms in episode s with respect to arm a_s^g can be bounded as follows:

$$\begin{aligned}
\mathbb{E}[R_{N,s}^{(1)}] &:= \mathbb{E}\left[\sum_{n\in\mathcal{N}_s} \mathbb{I}\{A_n \in \mathrm{BAD}_n\}(g_n(a_s^g) - g_n(A_n)) \right] \\
&= \mathbb{E}\left[\sum_{n\in\mathcal{N}_s} \sum_{a\in\mathrm{BAD}_n} \mathbb{I}\{A_n = a\}(g_n(a_s^g) - g_n(a)) \right] \\
&= \mathbb{E}\left[\sum_{n\in\mathcal{N}_s} \sum_{a\in\mathrm{BAD}_n} \sum_{\varepsilon\in\mathcal{B}} \mathbb{I}\{\mathrm{EXP}(a,n,\varepsilon)\} \sum_{t=n}^{\tau_n(\varepsilon,a)} \mathbb{I}\{A_t = a\}(g_t(a_s^g) - g_t(a)) \right] \\
&\le \mathbb{E}\left[\sum_{n\in\mathcal{N}_s} \sum_{a\in\mathrm{BAD}_n} \sum_{\varepsilon\in\mathcal{B}} \mathbb{I}\{\mathrm{EXP}(a,n,\varepsilon)\}\mathbb{I}\{\mathcal{E}_2 \cap \mathcal{E}_4\} \sum_{t=n}^{\tau_n(\varepsilon,a)} \mathbb{I}\{A_t = a\}(g_t(a_s^g) - g_t(a)) \right] \\
&\qquad + \mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_2 \cup \overline{\mathcal{E}}_4\} |\mathcal{N}_s| \right] \\
&\le \mathbb{E}\left[\sum_{n\in\mathcal{N}_s} \sum_{a\in\mathrm{BAD}_n} \sum_{\varepsilon\in\mathcal{B}} \mathbb{I}\{\mathrm{EXP}(a,n,\varepsilon)\} \left(34C_N \left(\frac{1}{\varepsilon} \vee C_N \right) + 20C_N^2 + 1 \right) \right] \\
&\qquad + \mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_2 \cup \overline{\mathcal{E}}_4\} |\mathcal{N}_s| \right]\,,
\end{aligned}$$

where the first inequality holds trivially when introducing the indicator for $\mathcal{E}_2 \cap \mathcal{E}_4$ as the sum is always at most $|\mathcal{N}_s|$, and the second inequality holds by (12). Taking into account that the conditional probability of $\mathrm{EXP}(a, n, \varepsilon)$ for $n \in \mathcal{N}_s$, given all the information up to the beginning of time step n, which also determines if $n \in \mathcal{N}_s$, is at most $\varepsilon/\sqrt{K(n + 1 - t_s)}$, the first term above can be bounded as

$$\mathbb{E}\left[\sum_{n\in\mathcal{N}_s} \sum_{a\in\mathrm{BAD}_n} \sum_{\varepsilon\in\mathcal{B}} \mathbb{I}\{\mathrm{EXP}(a,n,\varepsilon)\} \left(34C_N \left(\frac{1}{\varepsilon} \vee C_N \right) + 20C_N^2 + 1 \right) \right]$$

$$
\begin{aligned}
&\le \mathbb{E}\left[\sum_{n\in\mathcal{N}_s}\sum_{a\in\mathrm{BAD}_n}\sum_{\varepsilon\in\mathcal{B}}\frac{1}{\sqrt{K(n+1-t_s)}}\big(34C_N\left(1\vee\varepsilon C_N\right)+\varepsilon(20C_N^2+1)\big)\right]\\
&\le \mathbb{E}\left[\sum_{n\in\mathcal{N}_s}\sum_{a\in\mathrm{BAD}_n}\frac{1}{\sqrt{K(n+1-t_s)}\big(54C_N^2+34C_N(\log_2(N)+1)+1\big)}\right]\\
&\le \mathbb{E}\left[\sum_{n\in\mathcal{N}_s}\sum_{a\in\mathrm{BAD}_n}\frac{80C_N(\log_2(N)+1)}{\sqrt{K(n+1-t_s)}}\right]\\
&\le 160C_N(\log_2(N)+1)\mathbb{E}[\sqrt{K|\mathcal{N}_s|}],
\end{aligned}
$$

where in the second inequality we used that $|\mathcal{B}|\le\log_2(N)+1$ and that $\sum_{\varepsilon\in\mathcal{B}}\varepsilon<1$, and in the third one that for $K\le N^2$, $N\ge 64$ and $\delta=1/(\sqrt{K}N^{3/2})$, $54C_N^2+1\le 46C_N(\log_2(N)+1)$, while the last step follows by $\sum_{x=1}^{X}1/\sqrt{x}\le 2\sqrt{X}$. Combining with the above gives

$$
\mathbb{E}[R_{N,s}^{(1)}]\le 160C_N(\log_2(N)+1)\mathbb{E}[\sqrt{K|\mathcal{N}_s|}]+\mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_2\cup\overline{\mathcal{E}}_4\}|\mathcal{N}_s|\right]
$$

Summing up for all s we obtain

$$
\mathbb{E}[R_N^{(1)}]=\mathbb{E}\left[\sum_s R_{N,s}^{(1)}\right]\quad\le 160C_N(\log_2(N)+1)\mathbb{E}\left[\sum_s\sqrt{K|\mathcal{N}_s|}\right]+\mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_2\cup\overline{\mathcal{E}}_4\}\sum_s|\mathcal{N}_s|\right]\ .
$$

Now Lemma 3 and the choice of δ imply that $\mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_2\cup\overline{\mathcal{E}}_4\}\right]\le 2/(N\sqrt{KN})$, and hence the second term on the right hand side above can be bounded by $2/\sqrt{KN}$. To obtain the statements of the lemma, split the first term as

$$
\mathbb{E}\left[\sum_s\sqrt{K|\mathcal{N}_s|}\right]=\underbrace{\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1\cap\mathcal{E}_2\}\sum_s\sqrt{K|\mathcal{N}_s|}\right]}_{A}+\underbrace{\mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_1\cup\overline{\mathcal{E}}_2\}\sum_s\sqrt{K|\mathcal{N}_s|}\right]}_{B}\ .
$$

Since the number of episodes is at most $S+1$ under $\mathcal{E}_1\cap\mathcal{E}_2$ by Lemma 5, and since the $\mathcal{N}_s$ form a partition of $[1,N]$, we have $A\le\sqrt{K(S+1)N}$. To bound B, notice that $B\le N\sqrt{K}\mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_1\cup\overline{\mathcal{E}}_2\}\right]$. By Lemma 3, $\mathbb{E}\left[\mathbb{I}\{\overline{\mathcal{E}}_1\cup\overline{\mathcal{E}}_2\}\right]\le 2/(\sqrt{K}N^{3/2})$, and so $B\le 2/\sqrt{N}$. Putting everything together, we obtain

$$
\begin{aligned}
\mathbb{E}[R_N^{(1)}]&\le 160C_N(\log_2(N)+1)\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1\cap\mathcal{E}_2\}\sum_s\sqrt{K|\mathcal{N}_s|}\right]+\frac{320C_N(\log_2(N)+1)}{\sqrt{N}}+\frac{2}{\sqrt{KN}}\\
&\le 160C_N(\log_2(N)+1)\sqrt{K(S+1)N}+\frac{320C_N(\log_2(N)+1)}{\sqrt{N}}+\frac{2}{\sqrt{KN}},
\end{aligned}
$$

as desired. ■

Lemma 7 *Under the conditions of Theorem 1, we have*

$$
\begin{aligned}
\mathbb{E}\left[R_N^{(2)}\right]&\le 45C_N\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\}\sum_s\sqrt{4C_N^2K^2+K(t_{s+1}-t_s)}\right]+3/\sqrt{KN} \qquad (14)\\
&\le 46C_N\sqrt{K(S+1)N}+3/\sqrt{KN}\ . \qquad (15)
\end{aligned}
$$

Proof For any $a \in [K]$, let $\mathcal{N}_s(a)$ denote the set of time steps in episode s on which arm a is GOOD; note that by definition, $\mathcal{N}_s(a)$ is an interval of the form $[t_s, \tilde{t}_s(a)]$, where $\tilde{t}_s(a)$ denotes the last time step of episode s in which a is in the GOOD set. Since a^g_s belongs to GOOD throughout the whole episode, if $n \in \mathcal{N}_s(a)$, $P_n(a) = P_n(a^g_s)$, and we have

$$\begin{aligned}
&\mathbb{I}\{A_n = a\}(g_n(a^g_s) - g_n(a)) \\
&= g_n(a^g_s)\big(\mathbb{I}\{A_n = a\} - P_n(a)\big) + \big(P_n(a^g_s)g_n(a^g_s) - \mathbb{I}\{A_n = a^g_s\}r_t\big) \\
&\quad + \big(\mathbb{I}\{A_n = a^g_s\}r_t - \mathbb{I}\{A_n = a\}r_t\big) + \big(\mathbb{I}\{A_n = a\}r_t - P_n(a)g_n(a)\big) + g_n(a)\big(P_n(a) - \mathbb{I}\{A_n = a\}\big) .
\end{aligned}$$

Summing up for all $n \in \mathcal{N}_s(a)$, under $\mathcal{E}_1 \cap \mathcal{E}_3$ we get

$$\begin{aligned}
&\sum_{n\in\mathcal{N}_s(a)} \mathbb{I}\{A_n = a\}(g_n(a^g_s) - g_n(a)) \\
&\le \left(G'_{t_s:\tilde{t}_s(a)}(a, a^g_s) - G_{t_s:\tilde{t}_s(a)}(a, a^g_s)\right) + \left(G_{t_s:\tilde{t}_s(a)}(a^g_s) - \widehat{G}_{t_s:\tilde{t}_s(a)}(a^g_s)\right) \\
&\quad + \widehat{\Delta}_{t_s:\tilde{t}_s(a)}(a^g_s, a) + \left(\widehat{G}_{t_s:\tilde{t}_s(a)}(a) - G_{t_s:\tilde{t}_s(a)}(a)\right) + \left(G_{t_s:\tilde{t}_s(a)}(a, a) - G'_{t_s:\tilde{t}_s(a)}(a, a)\right) \\
&\le \widehat{\Delta}_{t_s:\tilde{t}_s(a)}(a^g_s, a) + 32C_N\left(\sqrt{P_{t_s:\tilde{t}_s(a)}(a)} \vee 2C_N\right) \qquad (16)
\end{aligned}$$

From here the regret of playing arms in the GOOD set with respect to a^g_s in episode s can be bounded as follows:

$$\begin{aligned}
\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\}R^{(2)}_{N,s} &:= \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \sum_{n\in\mathcal{N}_s} \mathbb{I}\{A_n \in \mathrm{GOOD}_n\}\big(g_n(a^g_s) - g_n(A_n)\big) \\
&= \sum_{a\in[K]} \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \sum_{n\in\mathcal{N}_s} \mathbb{I}\{A_n = a, a \in \mathrm{GOOD}_n\}\big(g_n(a^g_s) - g_n(a)\big) \\
&= \sum_{a\in[K]} \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \sum_{n\in\mathcal{N}_s(a)} \mathbb{I}\{A_n = a\}\big(g_n(a^g_s) - g_n(a)\big) \\
&\le \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \cdot \sum_{a\in[K]} \left(\widehat{\Delta}_{t_s:\tilde{t}_s(a)}(a^g_s, a) + 32C_N\left(\sqrt{P_{t_s:\tilde{t}_s(a)}(a)} \vee 2C_N\right)\right) \\
&\le \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \cdot \sum_{a\in[K]} \left(44C_N\left(\sqrt{P_{t_s:\tilde{t}_s(a)}(a)} \vee 2C_N\right) + 1\right) \\
&\le \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \cdot \sum_{a\in[K]} \left(44C_N\sqrt{4C_N^2 + P_{t_s:\tilde{t}_s(a)}(a)} + 1\right) \\
&\le \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \cdot 45C_N \sum_{a\in[K]} \sqrt{4C_N^2 + P_{t_s:t_{s+1}-1}(a)} \\
&\le \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \cdot 45C_N \sqrt{4C_N^2K^2 + K\sum_{a\in[K]}\sum_{t\in\mathcal{N}_s} P_t(a)} \\
&= \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_3\} \cdot 45C_N \sqrt{4C_N^2K^2 + K|\mathcal{N}_s|} ,
\end{aligned}$$

where the first inequality holds by (16), the second because the elimination condition does not hold at time step $\tilde{t}_s(a) - 1$ and the last term in the summation of $\widehat{\Delta}_{t_s:\tilde{t}_s(a)}(a^g_s, a)$ is at most 1, and the penultimate step follows from the Cauchy-Schwartz inequality.

Since we have at most $S+1$ episodes under $\mathcal{E}_1 \cap \mathcal{E}_2$ (by Lemma 5) and since $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ holds with probability at least $1 - 3/(N\sqrt{KN})$ (by Lemma 3 and the choice $\delta = 1/(\sqrt{K}N^{3/2})$), we obtain

$$\begin{aligned}\mathbb{E}\left[R_N^{(2)}\right] &\le \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3\} R_N^{(2)}\right] + 3/\sqrt{KN} \\ &\le \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3\} \sum_s 45 C_N \sqrt{4C_N^2 K^2 + K|\mathcal{N}_s|}\right] + 3/\sqrt{KN} \qquad (17) \\ &\le 45 C_N \sqrt{4C_N^2 K^2 (S+1)^2 + K(S+1)N} + 3\,, \qquad (18)\end{aligned}$$

where the last step follows again by the Cauchy-Schwartz inequality. Bounding the indicator by 1 in (17) proves the first inequality of the lemma.

To finish the proof, notice that if $180 C_N^2 K(S+1) \le N$, (18) implies

$$\mathbb{E}\left[R_N^{(2)}\right] \le \sqrt{45 \cdot 46}\, C_N \sqrt{K(S+1)N} + 3 \le 46 C_N \sqrt{K(S+1)N} + 3/\sqrt{KN}\,. \qquad (19)$$

Finally, if $180C_N^2 K(S+1) > N$, the right hand side above is lower bounded by $\frac{46}{\sqrt{180}} N > N$, hence (19) also holds trivially in this case (as the rewards are $[0,1]$-valued, and so $R_N^{(2)} \le N$). ■

4.4 Bounding $R_N^{(3)}$, the regret of playing arms a_s^g with respect to the optimal arms

In this section we bound the regret of playing arms a_s^g with respect to playing the optimal arms.

Lemma 8 *Let $\tau_1 = 1 < \tau_2 < \cdots < \tau_M \le \tau_{M+1} := N$ denote the time steps when either a new episode starts or the identity of the optimal arm changes.*[5] *Under the conditions of Theorem 1,*

$$\begin{aligned}\mathbb{E}[R_N^{(3)}] \le 25 C_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{i=1}^{M} \left(\left(21 + 2\sqrt{\log_2(N)+1}\right)\sqrt{K(\tau_{i+1} - \tau_i)} + 3C_N K\right)\right] \\ + 400 C_N.\end{aligned}$$

Furthermore, under the same conditions,

$$\mathbb{E}[R_N^{(3)}] \le 25 C_N \left(22 + 2\sqrt{\log_2(N)+1}\right)\sqrt{K(2S+1)N} + 400 C_N.$$

The rest of the section is devoted to proving this lemma. We introduce a partitioning of the time horizon into several intervals, and we bound the regret of a_s^g with respect to a_n^* (where $n \in \mathcal{N}_s$; recall that $\mathcal{N}_s$ is the set of time steps in episode s). Throughout we let $c_n(a) = g_n(a_n^*) - g_n(a)$ denote the instantaneous regret of an arm $a \in [K]$ in time step n, and $s(n)$ the index of the episode n belongs to (note that $s(n)$ is a random quantity).

[5] Note that M is random, and since by Lemma 5 the number of episodes is at most $S+1$ under $\mathcal{E}_1 \cap \mathcal{E}_2$, and the identity of the optimal arm changes S times, we have $M \le 2S+1$ when $\mathcal{E}_1 \cap \mathcal{E}_2$ holds (i.e., with high probability).

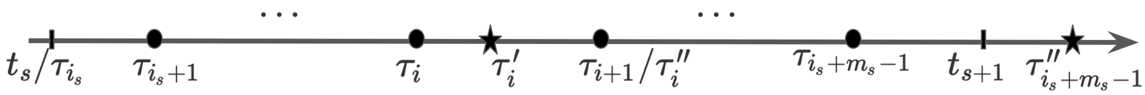

Figure 1: Partitioning the time horizon.

Partitioning the time horizon. In the analysis we partition the time horizon as follows: Each episode s is of the form $[\tau_{i_s} : \tau_{i_s+m_s} - 1]$ where i_s is the index marking the beginning of the episode and m_s is the (random) number of segment $[\tau_i : \tau_{i+1} - 1]$ in episode s. For any $i \in [M]$, let τ'_i be the first time step when a^*_t is not in the GOOD set, that is, $\tau'_i = \min\{t \in [\tau_i : \tau_{i+1} - 1] : a^*_t \in \text{BAD}_t\}$ if the latter set is non-empty, and we let $\tau'_i = \tau_{i+1}$ if the optimal arm is in the GOOD set during the entire segment $[\tau_i : \tau_{i+1} - 1]$. In what follows, we bound the regret on intervals $\{[\tau_i : \tau'_i - 1], [\tau'_i : \tau_{i+1} - 1] : i \in [M]\}$. Note that a^*_n and $a^g_{s(n)}$ are constant in the time steps n of any of these intervals. We also define τ''_i as the next point starting from τ'_i where the identity of the optimal arm changes, that is, $\tau''_i := \min\{\tau_j : a^*_{\tau_j} \neq a^*_{\tau_i}, j \in [i+1 : M]\}$ if this set is non-empty, and we define $\tau''_i := \tau_{M+1} = N$ otherwise. It is easy to see that under the event $\mathcal{E}_1 \cap \mathcal{E}_2$, τ''_i is either τ_{i+1} if τ_i corresponds to the start of an episode, and it is either τ_{i+1} or τ_{i+2} if it corresponds to a change of the optimal arm (see also Lemma 5). See Figure 1 for a visual depiction of the partitioning. For technical reasons, it will be advantageous to consider the regret on $[\tau'_i : \tau''_i - 1]$ instead of $[\tau'_i : \tau_{i+1} - 1]$, as for the former, the endpoint of the interval is determined by τ'_i, while it can be random for the latter even given τ'_i.

We start by analyzing time steps where the optimal arm is in the GOOD set in Section 4.4.1, and consider the significantly more complicated case when it belongs to the BAD set in Section 4.4.2. Some technical lemmas are presented in Section 4.6.

We start by defining a partitioning of the time horizon.

4.4.1 The optimal arm belongs to the GOOD set

In this section we analyze the regret on segments where a^g_s belongs to the set of GOOD arms, that is, for some $i \in [1, M]$, consider an interval $I = [\tau_i : \tau'_i - 1]$. Let $s = s(\tau_i)$ denote the episode I belongs to. Since both a^g_s and the optimal arm $a^*_I := a^*_n$ are constant over I and belong to GOOD_n for all $n \in I$, Lemma 4(ii) implies that if $\mathcal{E}_2$ holds, then

$$\sum_{n \in [\tau_i : \tau'_i - 2]} c_n(a^g_s) = K\widetilde{\Delta}_{\tau_i, \tau'_i - 2}(a^*_I, a^g_s) \leq 24C_N \left(\sqrt{K(\tau'_i - \tau_i)} \vee KC_N \right) .$$

Taking into account that $c_t \leq 1 \leq KC_N^2$, we have that under $\mathcal{E}_2$,

$$\sum_{n \in [\tau_i : \tau'_i - 1]} c_n(a^g_s) \leq 24C_N\sqrt{K(\tau'_i - \tau_i)} + 25KC_N^2 . \tag{20}$$

4.4.2 The optimal arm belongs to the BAD set

Fix an episode s (recall that this episode is $[\tau_{i_s} : \tau_{i_s+m_s} - 1]$). In what follows, we bound the regret of arm a^g_s for time steps when the optimal arm belongs to the BAD set, that is, over all intervals of the form $[\tau'_i, \tau_{i+1}-1]$ for $i \in [i_s : i_s + m_s - 1]$. Since a^g_s is determined only at the end of segment s, we bound the maximum regret caused by every arm $a \in [K]$ in this segment before it can be eliminated from the GOOD set. Furthermore, to avoid measurability problems, we bound this regret not on the original interval $[\tau'_i, \tau_{i+1}-1]$ (whose endpoint can be random given $\tau_{i+1}-1$) but on the possibly longer interval $[\tau'_i, \tau''_i - 1]$, whose endpoint is determined by τ'_i; importantly, throughout this subsection we do not consider that the episode can end earlier (at $\tau_{i+1}-1$) as this can only increase the regret and hence it gives an upper bound as desired. During the proof we consider the shortest intervals which ensure that, under $\mathcal{E}_2$, if the optimal arm is explored, arm a is eliminated by the optimal arm (see Lemma 4(ii)). The regret on these intervals is well-controlled, hence their overall contribution to $\mathbb{E}[R^{(3)}]$ is acceptable if there are not too many of them. On the other hand, the more such intervals are in the episode, the larger the probability of detecting the suboptimality of a, hence limiting a's contribution to the total regret.

Fix an arm a. Consider an index $i \in [i_s : i_s + m_s - 1]$ and the interval $[\tau'_i : \tau''_i - 1]$. For a time step $n \in [\tau'_i, \tau''_i - 1]$, let $n' \in [n : \tau''_i]$ be the smallest integer such that the following condition is satisfied for some $n'' \in [n : n']$:

$$\sum_{t=n''}^{n'-1} c_t(a) > 24 C_N \left(\sqrt{K(n'-n'')} \vee K C_N \right) . \tag{21}$$

If no such n' exists, we define $n' = \tau''_i$. Let $I_n := [n : n'-1]$. Note that if the optimal arm does not change in I_n, then the left-hand side of (21) is $K\widetilde{\Delta}_{n'':n'}(a^*_{n'}, a)$, and (21) is the elimination condition in Lemma 4(ii) (if both a and a^*_n are active in I_n). As such, an interval constructed in this way is called a *candidate* interval for eliminating a.

Given that $c_t \in [0,1]$ and $C_N \geq 1$, (21) implies $|I_n| = n' - n \geq n' - n'' \geq K C_N^2$, and thus $\sqrt{K|I_n|} \geq K C_N$. Therefore, since (21) does not hold for the interval $[n : n'-2]$ by definition and since $c_t(a) \leq C_N$ (as $c_t(a) \in [0,1]$ and $C_N \geq 1$),

$$\sum_{t \in I_n} c_t(a) \leq \begin{cases} 25 C_N \left(\sqrt{K|I_n|} \vee K C_N \right) = 25 C_N \sqrt{K|I_n|} & \text{if (21) holds for } n'; \\ 24 C_N \sqrt{K|I_n|} + 24 C_N^2 K & \text{otherwise.} \end{cases} \tag{22}$$

which provides a bound on the regret of a in I_n with respect to the optimal arm.

Let E_n denote the event that the optimal arm is active in interval I_n. By Lemma 4(ii), E_n leads to the elimination of a when $\mathcal{E}_2$ holds, hence we refer to E_n as an *elimination event.* If the optimal arm a^*_n is in BAD_n, it is active in I_n if a new exploration obligation of length at least $|I_n|/K$ is prescribed at the beginning of time step n.[6] Conditioned on the history, this happens with probability at least $\frac{1}{2\sqrt{|I_n|(n+1-t_s)}}$ since a new exploration obligation of length $1/\varepsilon^2$ is prescribed with probability $\varepsilon/\sqrt{K(n+1-t_s)}$ (see line 5 in Algorithm 1) for all $\varepsilon \in \mathcal{B}$ (the factor $\frac{1}{2}$ is due to the fact that ε can take values only in the exponential grid $\mathcal{B}$). Therefore,[7]

$$\mathbb{P}(\overline{E_n} | a^*_n \in \mathrm{BAD}_n, \mathcal{H}'_n) = \mathbb{P}(\overline{E_n} | a^*_n \in \mathrm{BAD}_n, \mathcal{H}'_{\tau'_i}) \leq 1 - \frac{1}{2\sqrt{|I_n|(n+1-t_s)}}. \tag{23}$$

[6] Recall that arm a is active in $[n' : n]$ if and only if $\mathrm{Active}_n(a) \leq n'$.

[7] Recall that history $\mathcal{H}'_n$ is defined in line 3 of Algorithm 1.

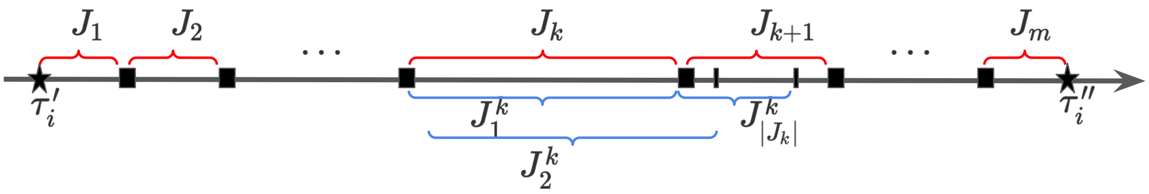

Figure 2: Partitioning of $[\tau_i', \tau_i'' - 1]$.

Let $\mathcal{J} = \{J_1, J_2, \ldots, J_m\}$ be a partitioning of $[\tau_i', \tau_i''-1]$ into candidate intervals ordered by their starting points (if $j < k$ then the starting point of J_j is smaller than that of J_k); we have $J_1 = I_{\tau_i'}$ and $J_j = I_{\tau_i'+|J_1|+\cdots+|J_{j-1}|}$ for $j \in [2 : m]$. Note that m is a random quantity which depends on τ_i', the starting point of the segment. See Figure 2 for a visual depiction. By the construction of candidate intervals, if an elimination event E_n happens in an interval J_k (i.e., $n \in J_k$), the corresponding exploration of the optimal arm finishes by the end of J_{k+1}, and if $\mathcal{E}_2$ holds, it means that a is eliminated from the GOOD set by then (when $a = a_s^g$, this means that the episode also finishes).

Let $F_k = \bigcup_{n \in J_k} E_n$ denote the event that an elimination event happens in J_k, and $\overline{F}^k := \bigcap_{j=1}^k \overline{F_j}$ that no elimination event happens before J_k. By the argument above, if $\mathcal{E}_2$ holds, then $\overline{F}^{k-1} \cap F_k$ means that a is eliminated from the GOOD set by an elimination event happening in J_k, hence a can only be in the GOOD set until the end of J_{k+1}. Let

$$R^G(a, [\tau_i' : \tau_i'' - 1]) = \sum_{n=\tau_i'}^{\tau_i''-1} \mathbb{I}\{a \in \text{GOOD}_n\} c_n(a)$$

denote the regret of arm a in interval $[\tau_i' : \tau_i'' - 1]$ when it is in the GOOD set.[8]

Lemma 9 *We have that*

$$\begin{aligned}&\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a, [\tau_i' : \tau_i''-1]) \\ &\qquad \le 25 C_N \left(2\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \left(\sqrt{K(\tau_i'' - \tau_i)} + C_N K \right) + \sum_{k=1}^{m-2} \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^k\} \sqrt{K|J_k|} \right).\end{aligned}$$

Proof We write

$$\begin{aligned}&\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \frac{R^G(a, [\tau_i' : \tau_i'' - 1])}{25 C_N} \\ &\qquad \le \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap F_1\}(\sqrt{K|J_1|} + \sqrt{K|J_2|}) \\ &\qquad\quad + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F_1} \cap F_2\}(\sqrt{K|J_1|} + \sqrt{K|J_2|} + \sqrt{K|J_3|})\end{aligned}$$

[8] With a slight abuse of notation, here and in the rest of this subsection we use GOOD_n to denote the GOOD set for the modified algorithm which does not start a new episode before τ_i'', as discussed before. The same applies to other notation, such as BAD_n.

$$
\begin{aligned}
&\quad + \ldots \\
&\quad + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F_1} \cap \ldots \cap \overline{F_{m-2}} \cap F_{m-1}\}(\sqrt{K|J_1|} + \ldots + \sqrt{K|J_m|} + C_N K) \\
&\quad + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F_1} \cap \ldots \cap \overline{F_{m-2}} \cap \overline{F_{m-1}}\}(\sqrt{K|J_1|} + \ldots + \sqrt{K|J_m|} + C_N K) \\
&\leq \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F_1} \cap F_2\}\sqrt{K|J_1|} \\
&\quad + \ldots \\
&\quad + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F_1} \cap \ldots \cap \overline{F_{m-2}} \cap F_{m-1}\}(\sqrt{K|J_1|} + \ldots + \sqrt{K|J_{m-2}|}) \\
&\quad + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F_1} \cap \ldots \cap \overline{F_{m-2}} \cap \overline{F_{m-1}}\}(\sqrt{K|J_1|} + \ldots + \sqrt{K|J_{m-2}|}) \\
&\quad + 2\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\left(\sqrt{K(\tau_i'' - \tau_i)} + C_N K\right) \\
&= \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^1\}\sqrt{K|J_1|} + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^2\}\sqrt{K|J_2|} \\
&\quad + \ldots + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^{m-2}\}\sqrt{K|J_{m-2}|} + 2\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\left(\sqrt{K(\tau_i'' - \tau_i)} + C_N K\right),
\end{aligned}
$$

where (i) the first inequality holds by the above argument about the elimination of a from the GOOD set and by (22) bounding its regret in intervals $J_1, J_2, \ldots$; (ii) the second inequality holds by bounding the last two terms in every row using $|J_k| \leq \tau_i'' - \tau_i$ for all k; (iii) the last equality follows by collecting like terms. Therefore,

$$
\begin{aligned}
&\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a, [\tau_i' : \tau_i'' - 1]) \\
&\quad \leq 25 C_N \left(2\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\left(\sqrt{K(\tau_i'' - \tau_i)} + C_N K\right) + \sum_{k=1}^{m-2} \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^k\}\sqrt{K|J_k|}\right).
\end{aligned}
$$

■

To proceed from here, we need to bound the probability of events $\overline{F}^k$ for $k \in [m-2]$. Let $\mathcal{J}^d = \{J_k \subset \mathcal{J} : |J_{k+1}| > \sum_{j=1}^k |J_j|\}$ denote the set of intervals where adding J_{k+1} at least doubles the length of the interval $\bigcup_{j=1}^k J_k$ covered so far (these are the intervals where the next lemma does not apply). We bound the regret of a on intervals in $\mathcal{J}^d$ separately. Because of the aforementioned doubling of length, and since $\left|\bigcup_{J \in \mathcal{J}} J\right| \leq \tau_i'' - \tau_i' \leq N$, $|\mathcal{J}^d| \leq \log_2(N) + 1$. Therefore, by (22) and the Cauchy-Schwartz inequality,

$$
\sum_{J \in \mathcal{J}^d} \sqrt{K|J|} \leq \sqrt{K(\tau_i'' - \tau_i')(\log_2(N) + 1)}. \tag{24}
$$

Next we bound the probability of events $\overline{F}^k$ on the remaining intervals.

Lemma 10 *Let $\mathcal{J}' = \mathcal{J} \setminus \mathcal{J}^d$. For any k such that $J_k \in \mathcal{J}'$,*

$$
\mathbb{P}\left(\overline{F}^k \middle| \mathcal{H}'_{\tau_i'}\right) \leq \exp\left(-\frac{1}{8\sqrt{\tau_i'' - t_s}} \sum_{j=1}^k \sqrt{|J_j|}\right).
$$

Proof Define $Q_I := \frac{1}{2\sqrt{|I|(\tau_i''-t_s)}}$ for any interval $I \subset [\tau_i', \tau_i''-1]$, and recall from (23) that for any $n \in [\tau_i' : \tau_i''-1]$,

$$\mathbb{P}(\overline{E_n} | a_n^* \in \mathrm{BAD}_n, \mathcal{H}'_{\tau_i'}) \le 1 - \frac{1}{2\sqrt{|I_n|(\tau_i''-t_s)}} = 1 - Q_{I_n}, . \tag{25}$$

Let $\{J_1^k, J_2^k, \ldots, J_{|J_k|}^k\}$ be the set of all candidate intervals starting in J_k (meaning that for $1 \le i \le |J_k|$, $J_i^k = I_t$ for some $t \in J_k$). See Figure 2 for a visual depiction. Note that by definition, all intervals J_i^k end no later than the last time step of J_{k+1}, hence $J_i^k \subset J_k \cup J_{k+1}$. Therefore, conditioned on τ_i', the probability that no elimination event happens before the end of J_k can be upper bounded as

$$\begin{aligned}
\mathbb{P}\left(\overline{F}^k \middle| \mathcal{H}'_{\tau_i'}\right) &= \mathbb{P}\Bigg(\bigcap_{n \in \cup_{j=1}^k J_j} \overline{E_n} \Bigg| \mathcal{H}'_{\tau_i'}\Bigg) \\
&= \prod_{n \in \cup_{j=1}^k J_j} \mathbb{P}(\overline{E_n} | \mathcal{H}'_{\tau_i'}, \overline{E_{\tau_i'}}, \ldots, \overline{E_{n-1}}) \\
&= \left(\prod_{n \in J_1} \mathbb{P}(\overline{E_n} | \mathcal{H}'_{\tau_i'}, \overline{E_{\tau_i'}}, \ldots, \overline{E_{n-1}})\right)\left(\prod_{n \in J_2} \mathbb{P}(\overline{E_n} | \mathcal{H}'_{\tau_i'}, \overline{E_{\tau_i'}}, \ldots, \overline{E_{n-1}})\right) \\
&\quad \cdots \left(\prod_{n \in J_k} \mathbb{P}(\overline{E_n} | \mathcal{H}'_{\tau_i'}, \overline{E_{\tau_i'}}, \ldots, \overline{E_{n-1}})\right) \\
&\le \left((1 - Q_{J_1^1}) \cdot (1 - Q_{J_2^1}) \ldots (1 - Q_{J_{|J_1|}^1})\right) \\
&\quad \cdot \left((1 - Q_{J_1^2}) \cdot (1 - Q_{J_2^2}) \ldots (1 - Q_{J_{|J_2|}^2})\right) \\
&\quad \cdots \\
&\quad \cdot \left((1 - Q_{J_1^k}) \cdot (1 - Q_{J_2^k}) \ldots (1 - Q_{J_{|J_k|}^k})\right),
\end{aligned}$$

where the last inequality holds by (25). We further upper bound the above probability by using the fact that $Q_I \ge Q_{I'}$ for any intervals $I \subset I'$ (and therefore $Q_{J_i^k} \ge Q_{J_k \cup J_{k+1}}$),

$$\begin{aligned}
\mathbb{P}\left(\overline{F}^k \middle| \mathcal{H}'_{\tau_i'}\right) &\le ((1 - Q_{J_1 \cup J_2}) \cdot (1 - Q_{J_1 \cup J_2}) \ldots (1 - Q_{J_1 \cup J_2})) \\
&\quad \cdot ((1 - Q_{J_2 \cup J_3}) \cdot (1 - Q_{J_2 \cup J_3}) \ldots (1 - Q_{J_2 \cup J_3})) \\
&\quad \cdots \\
&\quad \cdot \left((1 - Q_{J_k \cup J_{k+1}}) \cdot (1 - Q_{J_k \cup J_{k+1}}) \ldots (1 - Q_{J_k \cup J_{k+1}})\right) \\
&= (1 - Q_{J_1 \cup J_2})^{|J_1|} \cdot (1 - Q_{J_2 \cup J_3})^{|J_2|} \ldots (1 - Q_{J_k \cup J_{k+1}})^{|J_k|} \\
&\le \exp\left(-|J_1| Q_{J_1 \cup J_2}\right) \cdot \exp\left(-|J_2| Q_{J_2 \cup J_3}\right) \ldots \exp\left(-|J_k| Q_{J_k \cup J_{k+1}}\right) \\
&= \exp\left(-\frac{1}{2\sqrt{\tau_i'' - t_s}} \sum_{j=1}^k \frac{|J_j|}{\sqrt{|J_j| + |J_{j+1}|}}\right),
\end{aligned} \tag{26}$$

where in the last inequality we used that $1 - x \leq e^{-x}$ for any real number x. Then, by (26) and Lemma 13, for any k such that $J_k \in \mathcal{J}'$,

$$\begin{aligned}
\mathbb{P}\left(\overline{F}^k \middle| \mathcal{H}'_{\tau'_i}\right) &\leq \exp\left(-\frac{1}{2\sqrt{\tau''_i - t_s}} \sum_{j=1}^{k} \frac{|J_j|}{\sqrt{|J_j| + |J_{j+1}|}}\right) \\
&\leq \exp\left(-\frac{1}{8\sqrt{\tau''_i - t_s}} \sum_{j=1}^{k} \sqrt{|J_j|}\right) \\
&\leq \exp\left(\frac{1}{8\sqrt{\tau''_i - t_s}} \sum_{j \in [1:k]: J_j \in \mathcal{J}'} \sqrt{|J_j|}\right)
\end{aligned}$$

■

Now we are ready to bound the regret of an arm a in an episode s starting from time step τ_i when the optimal arm is in the BAD set and a belongs to the GOOD set, denoted as

$$R^G(a, \tau_i) := \sum_{j=i}^{i_s + m_s - 1} \sum_{n \in [\tau_j : \tau''_j - 1]} \mathbb{I}\{a \in \mathrm{GOOD}_n, a^*_n \in \mathrm{BAD}_n\} c_n(a)$$

Lemma 11 *Let* $D_N = 25 C_N \sqrt{K}$ *and* $D'_N = 10 + \sqrt{\log_2(N) + 1}$. *Furthermore, let* $\mathcal{S}_{\tau_i}$ *denote the collection of segments in episode* s *after* τ_i, *that is,* $\mathcal{S}_{\tau_i} = \{[\tau_j : \tau''_j - 1] : j \in [i : i_s + m_s - 1]\}$. *Then*

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a, \tau_i) \,\Big|\, \mathcal{H}'_{\tau_i}\Big] &\leq D_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{I \in \mathcal{S}_{\tau_i}} \left(D'_N \sqrt{|I|} + 2 C_N \sqrt{K}\right) \middle| \mathcal{H}'_{\tau_i}\right] \\
&\quad + 8 D_N \sqrt{\tau_i - t_s} + 8 D_N \mathbb{E}\left[(1 - \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}) \sum_{I \in \mathcal{S}_{\tau_i}} \sqrt{|I|} \middle| \mathcal{H}'_{\tau_i}\right]. \quad (27)
\end{aligned}$$

In particular, for $i = i_s$, *that is,* $\tau_{i_s} = t_s$, *we have*

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a, \tau_{i_s}) \,\Big|\, \mathcal{H}'_{\tau_{i_s}}\Big] \\
&\quad \leq D_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{I \in \mathcal{S}_{\tau_{i_s}}} \left(D'_N \sqrt{|I|} + 2 C_N \sqrt{K}\right) + 8\left(1 - \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\right) \sum_{I \in \mathcal{S}_{\tau_{i_s}}} \sqrt{|I|} \middle| \mathcal{H}'_{\tau_i}\right].
\end{aligned}$$

Proof In the proof we use the notation previously defined for the interval $[\tau'_i : \tau''_i - 1]$ without any reference to i, such as the set of intervals $\mathcal{J}$, the intervals $J_1, \ldots, J_m$ or the corresponding events $\overline{F}^k$. To simplify the notation, the latter will also be denoted by $\overline{F}^{J_k} := \overline{F}^k$.

We bound $\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a,\tau_i) | \mathcal{H}'_{\tau_i}\right]$ recursively as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a,\tau_i) | \mathcal{H}'_{\tau_i}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a,\tau_i) \,\middle|\, \mathcal{H}'_{\tau'_i}\right] \,\middle|\, \mathcal{H}'_{\tau_i}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a,[\tau'_i : \tau''_i - 1]) + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^{m-2}\} R^G(a,\tau''_i) \,\middle|\, \mathcal{H}'_{\tau'_i}\right] \,\middle|\, \mathcal{H}'_{\tau_i}\right] \\
&\le \mathbb{E}\Bigg[\mathbb{E}\Bigg[D_N \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\left((D'_N - 8)\sqrt{\tau''_i - \tau_i} + 2C_N\sqrt{K}\right) + D_N \sum_{J \in \mathcal{J}''} \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^J\}\sqrt{|J|} \\
&\qquad\qquad + \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \overline{F}^{m-2}\} R^G(a,\tau''_i) \,\Bigg|\, \mathcal{H}'_{\tau'_i}\Bigg] \,\Bigg|\, \mathcal{H}'_{\tau_i}\Bigg] \\
&\le D_N \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\left((D'_N - 8)\sqrt{\tau''_i - \tau_i} + 2C_N\sqrt{K}\right) + \sum_{J \in \mathcal{J}''} \mathbb{I}\{\overline{F}^J\}\sqrt{|J|} \,\middle|\, \mathcal{H}'_{\tau'_i}\right] \,\middle|\, \mathcal{H}'_{\tau_i}\right] \\
&\qquad + \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\overline{F}^{m-2}\} \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a,\tau''_i) \,\middle|\, \mathcal{H}'_{\tau''_i}\right] \,\middle|\, \mathcal{H}'_{\tau'_i}\right] \,\middle|\, \mathcal{H}'_{\tau_i}\right], \qquad (28)
\end{aligned}
$$

where (i) the second equality follows since $a \in \mathrm{GOOD}_{\tau''_i}$ is only possible (under $\mathcal{E}_2$) if $\mathcal{E}_2 \cap \overline{F}^{m-2}$ holds; (ii) the first inequality holds by Lemma 9 and (24); and (iii) the last inequality holds by dropping some of the indicators $\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}$ and because $\overline{F}^{m-2}$ is $\mathcal{H}'_{\tau''_i}$-measurable.

Now we are ready to prove (27) by induction. By definition, $R^G(a,\tau_k) = 0$ when a is eliminated from the GOOD set. This happens at latest at the end of the episode, so $R^G(a,\tau''_{i_s+m_s-1}) = 0$, satisfying (27) for $\tau''_{i_s+m_s-1}$ (note that we can define, without loss of generality $R^G(a,N+1) = 0$), so the starting assumption holds for backwards induction. Assume now that

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} R^G(a,\tau''_i) \,\middle|\, \mathcal{H}'_{\tau''_i}\right] &\le D_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{I \in \mathcal{S}_{\tau''_i}} \left(D'_N\sqrt{|I|} + 2C_N\sqrt{K}\right) \,\middle|\, \mathcal{H}'_{\tau''_i}\right] \\
&\qquad + 8D_N\sqrt{\tau''_i - t_s} + 8D_N \mathbb{E}\left[(1 - \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}) \sum_{I \in \mathcal{S}_{\tau''_i}} \sqrt{|I|} \,\middle|\, \mathcal{H}'_{\tau''_i}\right]
\end{aligned}
$$

holds. Combining this assumption with (28), it remains to prove that

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\sum_{J \in \mathcal{J}''} \mathbb{I}\{\overline{F}^J\}\sqrt{|J|} + 8\mathbb{I}\{\overline{F}^{m-2}\}\sqrt{\tau''_i - t_s} \,\middle|\, \mathcal{H}'_{\tau'_i}\right] \,\middle|\, \mathcal{H}'_{\tau_i}\right] \\
&\qquad\qquad \le 8\left(\sqrt{\tau_i - t_s} + \mathbb{E}\left[\sqrt{\tau''_i - \tau_i} \,\middle|\, \mathcal{H}'_{\tau_i}\right]\right) . \qquad (29)
\end{aligned}
$$

Let $m'' = |\mathcal{J}''|$, and let $x_1, \ldots, x_{m''}$ denote the length of the intervals $J \in \mathcal{J}''$ ordered according to the indices of the corresponding intervals (that is, if x_a is the length of $|J_k|$ and $x_{a'}$ is the length of $|J_{k'}|$, with $k < k'$, $J_k, J_{k'} \in \mathcal{J}''$, then $a < a'$). By the upper bound of Lemma 10 on $\mathbb{P}(\overline{F}^K | \mathcal{H}'_{\tau'_i})$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{J \in \mathcal{J}''} \mathbb{I}\{\overline{F}^J\} \sqrt{|J|} + 8\mathbb{I}\{\overline{F}^{m-2}\} \sqrt{\tau''_i - t_s} \,\middle|\, \mathcal{H}'_{\tau'_i} \right] \\
&\quad \le \sum_{k=1}^{m''} \exp\left(-\frac{1}{8\sqrt{\tau''_i - t_s}} \sum_{j=1}^{k} \sqrt{x_j} \right) \sqrt{x_k} + \exp\left(-\frac{1}{8\sqrt{\tau''_i - t_s}} \sum_{j=1}^{m''} \sqrt{x_j} \right) \cdot 8\sqrt{\tau''_i - t_s} \\
&\quad \le 8\sqrt{\tau''_i - t_s} \le 8\left(\sqrt{\tau_i - t_s} + \sqrt{\tau''_i - \tau_i} \right)
\end{aligned}$$

by Lemma 14 with $\alpha = 1/\left(8\sqrt{\tau''_i - t_s}\right)$ and $y_j = \sqrt{x_j}$. We get the desired result by taking conditional expectations $\mathbb{E}\left[\cdot \,\middle|\, \mathcal{H}'_{\tau_i}\right]$ of both sides. This shows that (29) holds, and hence so does the first statement of the lemma. The second statement holds because $\tau_{i_s} = t_s$. ■

Our desired regret bound follows easily from Lemma 11.

Proof [Proof of Lemma 8] Since a^g_s is in the GOOD set in the entire segment s,

$$R^G(a^g_s, \tau_{i_s}) = \sum_{i=i_s}^{i_s+m_s-1} \sum_{n \in [\tau'_i : \tau''_i - 1]} c_n(a^g_s)\,,$$

and so by the second part of Lemma 11,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{n=1}^{N} \mathbb{I}\{a^*_n \in \mathrm{BAD}_n\} c_n(a^g_{s(n)}) \right] \\
&\le D_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_s \sum_{I \in \mathcal{S}_{\tau_{i_s}}} \left(D'_N \sqrt{|I|} + 2C_N\sqrt{K} \right) + 8\left(1 - \mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\right) \sum_s \sum_{I \in \mathcal{S}_{\tau_{i_s}}} \sqrt{|I|} \right]. \\
&\le D_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{i=1}^{M} \left(D'_N \sqrt{\tau''_i - \tau_i} + 2C_N\sqrt{K} \right) \right] + 16 D_N \delta N^{3/2} \\
&\le 50 C_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{i=1}^{M} \left(\left(10 + \sqrt{\log_2(N) + 1}\right) \sqrt{K(\tau_{i+1} - \tau_i)} + C_N K \right) \right] + 400 C_N\,,
\end{aligned}$$

where the second inequality holds since $P(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - 2\delta = 1 - 2/(\sqrt{K}N^{3/2})$ by Lemma 3 (and by the choice of $\delta = 1/(\sqrt{K}N^{3/2})$) and since the number of intervals M in the second sum is at most N (also at most $2S+1$ by Lemma 5), and each has length at most N; and the last inequality follows by bounding $\sqrt{\tau''_i - \tau_i}$ with $\sqrt{\tau_{i+2} - \tau_{i+1}} + \sqrt{\tau_{i+1} - \tau_i}$ (which holds under $\mathcal{E}_1 \cap \mathcal{E}_2$ since then τ''_i is either τ_{i+1} or τ_{i+2}). Combining with (20) to cover the case when $a_n^* \in \mathrm{GOOD}_n$, and bounding $\tau_i' - \tau_i$ by $\tau_{i+1} - \tau_i$, we obtain

$$\mathbb{E}[R_N^{(3)}] \le 25C_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_{i=1}^{M} \left(\left(21 + 2\sqrt{\log_2(N)+1}\right)\sqrt{K(\tau_{i+1}-\tau_i)} + 3C_N K\right)\right] \\ + 400C_N.$$

By the Cauchy-Schwartz inequality it follows that

$$\begin{aligned}\mathbb{E}[R_N^{(3)}] &\le 25C_N \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \left(\left(21 + 2\sqrt{\log_2(N)+1}\right)\sqrt{KMN} + 3C_N KM\right)\right] + 400C_N \\ &\le 25C_N \left(21 + 2\sqrt{\log_2(N)+1}\right)\sqrt{K(2S+1)N} + 75C_N^2 K(2S+1) + 400C_N. \quad (30)\end{aligned}$$

Taking into account that the regret is at most N, as it was discussed in the proof of (19) from (18), it follows that the regret (30) is of order at most $C_N\sqrt{\log_2(N)}\sqrt{K(2S+1)N}$ (with $\delta = O(1/(\sqrt{K}N^{3/2}))$): If $N \ge 9C_N^2 K(2S+1)$, then

$$\mathbb{E}[R_N^{(3)}] \le 25C_N \left(22 + 2\sqrt{\log_2(N)+1}\right)\sqrt{K(2S+1)N} + 400C_N,$$

which bound also holds trivially if $N < 9C_N^2 K(2S+1)$, since in that case the right hand side is larger than N. This completes the proof of the lemma. ∎

4.5 Variational bound: proof of Corollary 2

In this section, we prove a variational bound on a slightly modified version of our algorithm, as described in Corollary 2. Putting together the first bounds in Lemmas 6-8, after straightforward manipulations we can obtain that

$$\mathbb{E}[R_N] \le \tilde{O}\left(\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\left(\sum_s \left(\sqrt{K(t_{s+1}-t_s)} \vee K\right) + \sum_i \left(\sqrt{K(\tau_{i+1}-\tau_i)} \vee K\right)\right)\right]\right) \quad (31)$$

where $\tilde{O}$ hides polylogarithmic factors.

For any $2 \le n' \le n'' \le N$, define

$$V_{n':n''-1} = \sum_{n=n'}^{n''-1} \max_{a\in[K]} |g_{n+1}(a) - g_n(a)|.$$

Note that $V = V_{1:N-1}$. We start by showing the following lemma:

Lemma 12 *Assume* ArmSwitch *is run with the modification described in Corollary 2 and that $KV \le N$.*[9] *Then*

$$\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_s (\sqrt{K(t_{s+1}-t_s)} \vee K) \le \sqrt{KN} + (1 + \sqrt[3]{2})(KV)^{1/3}N^{2/3}. \quad (32)$$

[9] Otherwise, our desired final $\tilde{O}((KV)^{1/3}N^{2/3})$ regret bound trivially holds.

Proof Fix an episode s, which is not the last one, and assume that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Since all arms are in the GOOD set at the beginning of step t_s, and all of them are eliminated by the end of $t_{s+1}-1$, there exists a sequence of different arms $a_1, \ldots, a_m \in [K]$ such that in episode s, a_k eliminates a_{k+1} for $k \in [m-1]$ and a_m eliminates a_1 (to simplify notation, we also denote $a_{m+1} := a_1$). This is easy to see by considering the directed graph whose nodes are the arms and there is an edge from arm a to arm b if a eliminates b: Since each arm is eliminated exactly once, there are K edges and each node has an in-degree one. Since there are K edges, there is an undirected circle in the graph, and since the in-degree within this circle is at most one for each node, the circle has to be a directed circle, and $a_1, \ldots, a_m$ can be chosen as the nodes in this circle such that a_1 is eliminated last from the GOOD set. Since a_m is eliminated before, a_m is already in the BAD set when it eliminates a_1.

For $k \in [m]$, consider the time step n when a_k eliminates the next arm a_{k+1}. If the elimination happens because of condition (7), the discussion after the condition implies that there is an interval $[n' : n] \subset [t_s : t_{s+1}-1]$ such that both a_k and a_{k+1} are active on this interval and $\Delta_{n':n}(a_k, a_{k+1}) > 0$, implying that there exists a time step $n_k \in [n' : n]$ such that $g_{n_k}(a_k) - g_{n_k}(a_{k+1}) > 0$. If the elimination happens because of the modified condition (8), it follows (as discussed in the paragraph following (6)) that $K\widetilde{\Delta}_{n':n}(a_k, a_{k+1}) > \sqrt{K(n-n'+1)}$ for some interval $[n' : n] \subset [t_s : t_{s+1}-1]$ where both arms are active. It follows then that there exists a time step $n_k \in [n' : n]$ such that $g_{n_k}(a_k) - g_{n_k}(a_{k+1}) > \sqrt{\frac{K}{n-n'+1}} \ge \sqrt{\frac{K}{t_{s+1}-t_s}}$. Note that if $a_k \in$ BAD when it eliminates a_{k+1}, only condition (8) can be triggered, and so we obtain

$$g_{n_k}(a_k) - g_{n_k}(a_{k+1}) \ge \begin{cases} \sqrt{\frac{K}{t_{s+1}-t_s}} & \text{if } a_k \text{ belongs to BAD when it eliminates } a_{k+1}; \\ 0 & \text{otherwise.} \end{cases}$$

Now clearly a_m is in the BAD set when it eliminates a_1 $(= a_m)$, hence

$$\begin{aligned}
\sqrt{\frac{K}{t_{s+1}-t_s}} &\le g_{n_1}(a_1) - g_{n_1}(a_2) + g_{n_2}(a_2) - g_{n_2}(a_3) + \cdots + g_{n_m}(a_m) - g_{n_m}(a_1) \\
&\le |g_{n_1}(a_1) - g_{n_m}(a_1)| + |g_{n_2}(a_2) - g_{n_1}(a_2)| + \cdots + |g_{n_m}(a_m) - g_{n_{m-1}}(a_m)| \\
&\le 2 \sum_{n=n_1}^{n_m-1} \max_{a \in [K]} |g_{n+1}(a) - g_n(a)| = 2V_{n_1:n_m-1} \\
&\le 2V_{t_s:t_{s+1}-1}.
\end{aligned}$$

Denoting the number of episodes by S' and bounding the contribution of the last interval as $\sqrt{K(t_{S'+1}-t_{S'})} \le \sqrt{KN}$, Hölder's inequality implies

$$\begin{aligned}
\sum_{s \in [S']} \sqrt{K(t_{s+1}-t_s)} &\le \sqrt{KN} + \left(\sum_{s \in [S'-1]} \sqrt{\frac{K}{t_{s+1}-t_s}} \right)^{1/3} \left(\sum_{s \in [S'-1]} (t_{s+1}-t_s)\sqrt{K} \right)^{2/3} \\
&\le \sqrt{KN} + \left(\sum_{s \in [S'-1]} 2V_{t_s:t_{s+1}-1} \right)^{1/3} K^{1/3} N^{2/3}
\end{aligned}$$

$$= \sqrt{KN} + (2KV)^{1/3}N^{2/3} .$$

On the other hand, by $V_{t_s:t_{s+1}-1} \geq \sqrt{K/(t_{s+1}-t_s)}$, the inequality $t_{s+1}-t_s \leq K$ implies $V_{t_s,t_{s+1}} \geq 1$. Therefore, using the fact that $KV \leq N$,

$$\sum_s K\mathbb{I}\{t_{s+1}-t_s \leq K\} \leq \sum_s KV_{t_s,t_{s+1}} \leq KV \leq \sqrt{KN} + (KV)^{1/3}N^{2/3} .$$

We conclude that under $\mathcal{E}_1 \cap \mathcal{E}_2$, $\sum_s \sqrt{K(t_{s+1}-t_s)} \vee K \leq (1+\sqrt[3]{2})(KV)^{1/3}N^{2/3})$, as desired. ■

It would be tempting to bound the second term in (31) similarly; however, the length of the intervals $\tau_{i+1}-\tau_i$ (typically between two time steps where the optimal arm changes) is not connected to the variations of the reward function. To resolve this problem, we can modify our Lemma 8 and consider time steps where the rewards change significantly, as defined by Suk and Kpotufe (2022).[10] We say that an arm a suffers significant regret in an interval $[n' : n]$, if $\sum_{t\in[n':n]} (g_t(a_t^*) - g_t(a)) \geq \sqrt{K(n-n'+1)}$. Then starting with $\sigma_1 = 1$, the time step σ_i for the $(i-1)$th significant change is defined as the first time step σ_i after σ_{i-1} such that for every arm a there exists an interval $I_a \subset [\sigma_{i-1} : \sigma_i]$ such that a suffers a significant regret in I_a. The last arm to suffer significant regret in $[\sigma_{i-1} : \sigma_i]$ (or one such arm with the largest cumulative reward in the interval if there are more than one), denoted by $\hat{a}_i^*$ and called the approximately optimal arm for $[\sigma_{i-1} : \sigma_i - 1]$, plays the role of the optimal arm in the interval $[\sigma_{i-1} : \sigma_i - 1]$ (by definition, its regret compared to the sequence of optimal arms in $[\sigma_i : \sigma_{i+1}-1]$ is bounded by $\sqrt{K(\sigma_i - \sigma_{i-1})}$). Note that the σ_i and the $\hat{a}_i^*$ are deterministic, as they only depend on reward functions $(g_1, \ldots, g_n)$.

It is easy to see that the proof of Lemma 8 goes through if we define τ_i with the time steps of significant changes instead of the changes in the optimal arm, with an extra term of $\sum_i \sqrt{K(\sigma_i - \sigma_{i-1})}$, accounting for using approximately optimal arms instead of the optimal arms in the analysis. Then the second term in (31) can be bounded by (for the redefined τ_i) as

$$\tilde{O}\left(\sum_s (\sqrt{K(t_{s+1}-t_s)} \vee K) + \sum_i (\sqrt{K(\sigma_{i+1}-\sigma_i)} \vee K)\right)$$

giving the overall bound

$$\mathbb{E}[R_N] \leq \tilde{O}\Bigg(\mathbb{E}\Bigg[\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\}\Bigg(\sum_s \left(\sqrt{K(t_{s+1}-t_s)} \vee K\right) + \sum_i \left(\sqrt{K(\sigma_{i+1}-\sigma_i)} \vee K\right)\Bigg)\Bigg]\Bigg) \tag{33}$$

As noticed by Suk and Kpotufe (2022), the second term above can also be bounded similarly to Lemma 12: It follows easily from the definition of significant change that for all but the last intervals $[\sigma_i : \sigma_{i+1}-1]$, there is a time step $n_i \in [\sigma_i : \sigma_{i+1}-1]$ such that $g_{n_i}(a^*_{n_i}) - g_{n_i}(a^*_{\sigma_{i+1}}) \geq \sqrt{K/(\sigma_{i+1}-\sigma_i)}$ (the largest gap in the interval where $a^*_{\sigma_{i+1}}$ suffers a significant regret – note that this interval finishes before σ_{i+1} because of the optimality of $a^*_{\sigma_{i+1}}$) and $g_{\sigma_{i+1}}(a^*_{\sigma_{i+1}}) - g_{\sigma_{i+1}}(a^*_{n_i}) \geq 0$. Therefore,

$$\sqrt{K/(\sigma_{i+1}-\sigma_i)} \leq g_{n_i}(a^*_{n_i}) - g_{n_i}(a^*_{\sigma_{i+1}}) + g_{\sigma_{i+1}}(a^*_{\sigma_{i+1}}) - g_{\sigma_{i+1}}(a^*_{n_i}) \leq 2V_{n_i:\sigma_{i+1}-1} \leq 2V_{\sigma_i:\sigma_{i+1}-1}.$$

[10] We present a slightly different definition than that of Suk and Kpotufe (2022).

Similarly to Lemma 12, this yields

$$\mathbb{I}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \sum_i (\sqrt{K(\sigma_i - \sigma_i)} \vee K) \le \sqrt{KN} + (1 + \sqrt[3]{2})(KV)^{1/3} N^{2/3} \tag{34}$$

when $KV \le N$.

Proof [Proof of Corollary 2] Combining Lemma 6, Lemma 7, and the modified version (33) of Lemma 8 with Lemma 12 and (34) proves the corollary. ■

4.6 Technical lemmas

Lemma 13 *Let $x_1, \ldots, x_{n+1}$ be a sequence of positive reals. Then*

$$\frac{1}{2} \sum_{j=1}^{n} \sqrt{x_j} - \frac{\sqrt{x_{n+1}}}{4} \le \sum_{j=1}^{n} \frac{x_j}{\sqrt{x_j + x_{j+1}}} \ .$$

In particular, if $x_{n+1} \le \sum_{j=1}^{n} x_j$, then

$$\frac{1}{4} \sum_{j=1}^{n} \sqrt{x_j} \le \sum_{j=1}^{n} \frac{x_j}{\sqrt{x_j + x_{j+1}}} \ .$$

Proof We have

$$\begin{aligned} \sum_{j=1}^{n} \frac{x_j}{\sqrt{x_j + x_{j+1}}} &\ge \sum_{j=1}^{n} \frac{x_j}{\sqrt{x_j} + \sqrt{x_{j+1}}} && (35) \\ &= \sum_{j=1}^{n} \left(\frac{x_j - x_{j+1}}{\sqrt{x_j} + \sqrt{x_{j+1}}} + \frac{x_{j+1}}{\sqrt{x_j} + \sqrt{x_{j+1}}} \right) \\ &= \sum_{j=1}^{n} \left(\sqrt{x_j} - \sqrt{x_{j+1}} + \frac{x_{j+1}}{\sqrt{x_j} + \sqrt{x_{j+1}}} \right) \\ &= \sqrt{x_1} - \sqrt{x_{n+1}} + \sum_{j=1}^{n} \frac{x_{j+1}}{\sqrt{x_j} + \sqrt{x_{j+1}}} \ . && (36) \end{aligned}$$

Combining (35) and (36), we obtain

$$\begin{aligned} \sum_{j=1}^{n} \frac{x_j}{\sqrt{x_j + x_{j+1}}} &\ge \frac{\sqrt{x_1} - \sqrt{x_{n+1}}}{2} + \frac{1}{2} \sum_{j=1}^{n} \frac{x_j + x_{j+1}}{\sqrt{x_j} + \sqrt{x_{j+1}}} \\ &\ge \frac{\sqrt{x_1} - \sqrt{x_{n+1}}}{2} + \frac{1}{4} \sum_{j=1}^{n} \left(\sqrt{x_j} + \sqrt{x_{j+1}} \right) \end{aligned}$$

$$\geq \frac{1}{2}\sum_{j=1}^{n}\sqrt{x_j} - \frac{\sqrt{x_{n+1}}}{4} ,$$

finishing the proof of the first statement. The second statement holds because $x_{n+1} \leq \sum_{j=1}^{n} x_j$ implies $\sqrt{x_{n+1}} \leq \sqrt{\sum_{j=1}^{n} x_j} \leq \sum_{j=1}^{n}\sqrt{x_j}$. ■

Lemma 14 *Let $\alpha, y_1, \ldots, y_m$ be positive real numbers. Then*

$$\sum_{k\in[m]} \exp\left(-\alpha\sum_{j=1}^{k} y_j\right) y_k + \exp\left(-\alpha\sum_{j=1}^{m} y_j\right)\frac{1}{\alpha} \leq \frac{1}{\alpha} . \tag{37}$$

Proof First, we show that $e^{-\alpha y}y + e^{-\alpha y}\frac{1}{\alpha} \leq \frac{1}{\alpha}$. Let $f(y) = e^{-\alpha y}y + e^{-\alpha y}\frac{1}{\alpha}$, and observe that

$$f'(y) = e^{-\alpha y} - \alpha e^{-\alpha y}y - e^{-\alpha y} = -\alpha e^{-\alpha y}y \leq 0 .$$

Therefore f is maximized at $y = 0$, and so $f(y) \leq f(0) = \frac{1}{\alpha}$. We finish the proof by induction. Assume (37) holds for $m-1$ (in place of m). Then

$$\begin{aligned}
&\sum_{k\in[m]} \exp\left(-\alpha\sum_{j=1}^{k} y_j\right) y_k + \exp\left(-\alpha\sum_{j=1}^{m} y_j\right)\frac{1}{\alpha}\\
&= \sum_{k\in[m-1]} \exp\left(-\alpha\sum_{j=1}^{k} y_j\right) y_k + \exp\left(-\alpha\sum_{j=1}^{m-1} y_j\right)\left(e^{-\alpha y_m}y_m + e^{-\alpha y_m}\frac{1}{\alpha}\right)\\
&\leq \sum_{k\in[m-1]} \exp\left(-\alpha\sum_{j=1}^{k} y_j\right) y_k + \exp\left(-\alpha\sum_{j=1}^{m-1} y_j\right)\frac{1}{\alpha}\\
&\leq \frac{1}{\alpha} ,
\end{aligned}$$

where in the last step we used the induction hypothesis. This completes the proof. ■

5. Conclusions and future directions

We have introduced ArmSwitch, an algorithm for learning in non-stationary stochastic multi-armed bandit environments. The main feature of our algorithm is that its regret scales as $\widetilde{O}(\sqrt{KSN})$, where S is the number of changes in the identity of the optimal arm, and it is unknown to the algorithm. In contrast, existing works for this problem bound the regret in terms of the number of changes in reward functions, which can be much larger.

An interesting related question is whether similar bounds can be shown on the regret with respect to any sequence of arms (instead of the sequence of optimal arms considered here). This can be achieved by the EXP3.S algorithm of Auer et al. (2002) if S, the number of changes in the sequence of reference arms in this case, is known and can be used to tune the method. However, obtaining such a result without the prior knowledge of S might require fundamentally different techniques, because we cannot eliminate arms which proved to be suboptimal at a given time step (as they can still remain in the reference sequence).

While we also provided a variational regret bound (which depends on the temporal variation of reward functions instead of the time horizon), this bound is unsatisfactory in the sense that it may be dominated by variations of the rewards of suboptimal arms. In fact, it is not clear which variations of the reward function are important, and exploring this problem is an interesting question for future research.

Another interesting future direction is designing algorithms with similar guarantees in the more complex setting of reinforcement learning. Unlike in the bandit setting, it is not clear what notion of complexity we should use in a reinforcement learning problem. A straightforward extension would be regret bounds that scale with the square root of the number of changes in the optimal policy multiplied by the number of policies. However, the space of policies is prohibitively large, which would make such bounds less meaningful. A more refined variant could weight each change by the probability of visiting a state where a change happens; we leave the exploration of these – and other – ideas for future work.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *ICML*, pages 1638–1646, 2014.

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 and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed 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 *European Workshop on Reinforcement Learning*, volume 14, page 375, 2018.

Peter Auer, Yifang Chen, Pratik Gajane, Chung-Wei Lee, Haipeng Luo, Ronald Ortner, and Chen-Yu Wei. Achieving optimal dynamic regret for non-stationary bandits without prior information. In *COLT*, 2019a.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *COLT*, 2019b.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27:199–207, 2014.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

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 *COLT*, 2019.

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

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

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. In *35th Annual Conference on Learning Theory*, 2022.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In *COLT*, 2021.

Appendix A. Auxiliary proofs

We start with a Freedman-style martingale tail inequality:

Theorem 15 (Agarwal et al., 2014) *Let $(\mathcal{H}_t; t \geq 1)$ be a filtration, $(X_t; t \geq 1)$ be a real-valued martingale difference sequence adapted to $(\mathcal{H}_t)$ (i.e., X_{t-1} is H_t-measurable and $\mathbb{E}[X_t|\mathcal{H}_t] = 0$). If $|X_t| \leq B$ almost surely, then for any $\eta \in (0, 1/B]$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{n} X_t \leq \eta(e-2) \sum_{t=1}^{n} \mathbb{E}[X_t^2|\mathcal{H}_t] + \frac{\log(1/\delta)}{\eta} \ .$$

The optimal η minimizing the above bound is $\eta^* = \sqrt{\frac{\log(1/\delta)}{(e-2)\sum_{t=1}^{n} \mathbb{E}[X_t^2|\mathcal{H}_{t-1}]}}$, which would lead to a bound

$$\sum_{t=1}^{n} X_t \leq 2\sqrt{(e-2)\log(1/\delta) \sum_{t=1}^{n} \mathbb{E}[X_t^2|\mathcal{H}_t]} \ .$$

However, since η^* depends on the sum of $\mathbb{E}[X_t^2|\mathcal{H}_t]$, it is not guaranteed that $\eta^* \leq 1/B$. The next corollary takes care of these problems.

Corollary 16 *Under the conditions of Theorem 15, for any* $0 < \delta \le 2/e$,

$$\left|\sum_{t=1}^{n} X_t\right| \le 2e\sqrt{(e-2)\log\left(\frac{\log(n/B)+2}{\delta}\right)\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_t]} \ \vee \ 2B\log\left(\frac{\log(n/B)+2}{\delta}\right)$$

holds with probability at least $1-\delta$.

Proof We first prove the bound on $\sum_{t=1}^{n} X_t$ with $\delta' = \delta/2$, then we apply the same bound to $\sum_{t=1}^{n}(-X_t)$. We create an exponential grid for η; applying Theorem 15 for each value in the grid together with the union bound will prove the corollary. Since $|X_t| \le B$, the smallest possible value of η^* is $\eta_{\min} = \sqrt{\frac{\log(1/\delta')}{(e-2)nB}}$. Assume first that $\eta_{\min} \le 1/B$. Let $\mathcal{G} = \{e^{-i}/B : i \in \lfloor \log(n/B)/2 \rfloor\}$. Then, by Theorem 15 and the union bound, with probability at least $1-\delta'$, we have

$$\sum_{t=1}^{n} X_t \le \eta(e-2)\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_t] + \frac{\log(\log(n/B)/2+1)/\delta')}{\eta} \tag{38}$$

for all $\eta \in \mathcal{G}$ simultaneously.

The smallest element of $\mathcal{G}$, denoted $\eta'_{\min}$, satisfies

$$\eta'_{\min} = e^{-\lfloor \log(n/B)/2 \rfloor}/B \le e/\sqrt{Bn} \le e\sqrt{\frac{\log(1/\delta')}{(e-2)nB}} = e\eta_{\min}$$

where the second inequality holds because $\delta' \le 1/e$. Therefore, for any $\eta \in [\eta_{\min}, 1/B]$ there is an $\eta' \in \mathcal{G}$ such that $\eta \le \eta' \le \eta e$. Thus, by (38), for any $\eta \in [\eta_{\min}, 1/B]$,

$$\begin{aligned}\sum_{t=1}^{n} X_t &\le \eta'(e-2)\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_{t-1}] + \frac{\log(\log(n/B)/2+1)/\delta')}{\eta'} \\ &\le e\left[\eta(e-2)\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_t] + \frac{\log(\log(n/B)/2+1)/\delta')}{\eta}\right].\end{aligned}$$

Specifically, for the minimizer $\eta_m = \sqrt{\frac{\log(\log(n/B)/2+1)/\delta')}{(e-2)\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_t]}}$ of the above bound, we obtain

$$\sum_{t=1}^{n} X_t \le 2e\sqrt{(e-2)\log\left(\frac{\log(n/B)/2+1}{\delta'}\right)\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_t]} \tag{39}$$

If $\eta_m \le 1/B$, the above bound holds. If not,

$$\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_t] \le B^2\frac{\log(\log(n/B)/2+1)/\delta')}{e-2}$$

and hence by (38) for $\eta = 1/B \in \mathcal{G}$, we get

$$\sum_{t=1}^{n} X_t \le 2B\log\left(\frac{\log(n/B)/2+1}{\delta'}\right).$$

Since one of the last two inequalities always hold, we obtain that with probability at least $1-\delta'$,

$$\sum_{t=1}^{n} X_t \le 2e\sqrt{(e-2)\log\left(\frac{\log(n/B)/2+1}{\delta'}\right)\sum_{t=1}^{n}\mathbb{E}[X_t^2|\mathcal{H}_t]} \;\vee\; 2B\log\left(\frac{\log(n/B)/2+1}{\delta'}\right) .$$

Using the same bound for $(-X_t)$, and using the union bound proves the corollary. ■

Now we are ready to prove Lemma 3.

Proof of Lemma 3. Proof for $\mathcal{E}_1$: Fix a time interval $[n':n] \subset [N]$ and an arm $a \in [K]$. For any $t \in [n':n]$, define variable $X_t = \mathbb{I}\{A_t = a\}r_t - P_t(a)g_t(a)$. Then $|X_t| \le 1$, $\mathbb{E}[X_t|\mathcal{H}_t] = 0$, and

$$\begin{aligned}\mathbb{E}[X_t^2|\mathcal{H}_t, A_t] &= \mathbb{E}\left[\Big(\mathbb{I}\{A_t = a\}\big(r_t - g_t(a)\big) + \big(\mathbb{I}\{A_t = a\} - P_t(a)\big)g_t(a)\Big)^2 \Big|\mathcal{H}_t, A_t\right] \\ &= \mathbb{I}\{A_t = a\}\mathbb{E}\left[\big(r_t - g_t(a)\big)^2|\mathcal{H}_t, A_t\right] + \big(\mathbb{I}\{A_t = a\} - P_t(a)\big)^2 g_t^2(a).\end{aligned}$$

Therefore, since r_t takes values in $[0,1]$, $\mathrm{Var}[r_t|A_t, \mathcal{H}_t] \le 1/4$,

$$\mathbb{E}[X_t^2|\mathcal{H}_t] \le P_t(a)/4 + P_t(a)(1 - P_t(a))g_t^2(a) \le 5P_t(a)/4.$$

Therefore, for any fixed $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\begin{aligned}\left|\sum_{t=n'}^{n} X_t\right| &\le 2e\sqrt{\frac{5}{4}(e-2)\log\left(\frac{\log(n-n'+1)+2}{\delta'}\right)P_{n':n}(a)} \vee 2\log\left(\frac{\log(n-n'+1)+2}{\delta'}\right) \\ &\le 6\sqrt{\log\left(\frac{\log(n-n'+1)+2}{\delta'}\right)P_{n':n}(a)} \vee 2\log\left(\frac{\log(n-n'+1)+2}{\delta'}\right)\end{aligned}$$

Using $\delta' = \delta/(2KN^2)$ and the union bound over all intervals $[n':n]$ and arms a (note that this choice of δ' satisfies $\delta' \le 2/e$ for all $\delta \in (0,1)$), we obtain that with probability at least $1-\delta$

$$\left|\widehat{G}_{n:n'}(a) - G_{n:n'}(a)\right| \le 6C_{n,n'}(\sqrt{P_{n:n'}(a)} \vee C_{n,n'}),$$

finishing the proof of the bound for $\mathcal{E}_1$.

For $\mathcal{E}_2$, for any arm $a \in [K]$ active in $[n':n]$, letting $X_t = \mathbb{I}\{\widetilde{A}_t = a\}r_t - g_t(a)/K$ yields the desired bound similarly, because here one can show that $\mathbb{E}[X_t^2|\mathcal{H}_t] \le 5/(4K)$.

For $\mathcal{E}_3$, let $X_t = (\mathbb{I}\{A_t = a\} - P_t(a))g_t(a')$. Then $\mathbb{E}[X_t^2|\mathcal{H}_t] \le P_t$, and we obtain the bound

$$\left|G'_{n:n'}(a,a') - G_{n:n'}(a,a')\right| \le 5C'_{n,n'}(\sqrt{P_{n:n'}(a)} \vee C'_{n,n'}),$$

where we introduced $C'_{n,n'}$ because here the union must be taken over all ordered pairs of arms $a, a' \in [K]$ instead of all arms.

Finally, for $\mathcal{E}_4$ introducing $X_t = \big(g_t(a) - g_t(a')\big)\big(\mathbb{I}\{\widetilde{A}_t = a\} - 1/K\big)$ for arms a, a' active in $[n' : n]$ gives $\mathbb{E}[X_t^2 | \mathcal{H}_t] \le 1/K$, which yields the desired bound. ■

Proof of Lemma 4. If $\Delta_{n':n}(a', a) > 24 C_{n',n} \left(\sqrt{P_{n':n}} \vee C_{n',n}\right)$ for $a, a' \in \mathrm{GOOD}_n$, then by the definition of $\mathcal{E}_1$ and by (9), $\textsc{Elim}_n(a', a)$ is true:

$$\widehat{\Delta}_{n':n}(a', a) \ge \Delta_{n':n}(a', a) - 12 C_{n',n} \left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right) > 12 C_{n',n} \left(\sqrt{P_{n':n}(a)} \vee C_{n',n}\right),$$

proving the first part of (i). If $a, a' \in \mathrm{GOOD}_{n+1}$, then a is not eliminated at the end of time step n, hence $\textsc{Elim}_n(a', a)$ is false. Consequently, $\Delta_{n':n}(a', a) \le 24 C_{n',n} \left(\sqrt{P_{n':n}} \vee C_{n',n}\right)$, finishing the proof of part (i). Part (ii) can be shown similarly. ■