Proceedings of Machine Learning Research vol 99:1–30, 2019 32nd Annual Conference on Learning Theory

A New Algorithm for Non-stationary Contextual Bandits: Efficient, Optimal, and Parameter-free

Yifang Chen YIFANG@USC.EDU
Chung-Wei Lee LEECHUNG@USC.EDU
Haipeng Luo HAIPENGL@USC.EDU
Chen-Yu Wei CHENYU.WEI@USC.EDU
University of Southern California

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We propose the first contextual bandit algorithm that is parameter-free, efficient, and optimal in terms of dynamic regret. Specifically, our algorithm achieves $\mathcal{O}(\min\{\sqrt{KST}, K^{\frac{1}{3}}\Delta^{\frac{1}{3}}T^{\frac{2}{3}}\})$ dynamic regret for a contextual bandit problem with T rounds, K actions, S switches and Δ total variation in data distributions. Importantly, our algorithm is adaptive and does not need to know S or Δ ahead of time, and can be implemented efficiently assuming access to an ERM oracle.

Our results strictly improve the $\mathcal{O}(\min\{S^{\frac{1}{4}}T^{\frac{3}{4}}, \Delta^{\frac{1}{5}}T^{\frac{4}{5}}\})$ bound of (Luo et al., 2018), and greatly generalize and improve the $\mathcal{O}(\sqrt{ST})$ result of (Auer et al., 2018) that holds only for the two-armed bandit problem without contextual information. The key novelty of our algorithm is to introduce *replay phases*, in which the algorithm acts according to its previous decisions for a certain amount of time in order to detect non-stationarity while maintaining a good balance between exploration and exploitation.

Keywords: contextual bandit, non-stationarity, optimal dynamic regret, oracle-efficiency, parameter-free, replay

1. Introduction

For online learning problems, a standard performance measure is *static regret*, which compares the total reward of the best fixed policy (or action/arm/expert under different contexts) and that of the algorithm. While minimizing static regret makes sense when there exists a fixed policy with large total reward, it becomes much less meaningful in a non-stationary environment where data distribution is changing over time and no single policy can perform well all the time.

Instead, in this case a more natural benchmark would be to compare the algorithm with the *best sequence of policies*. This is formally defined as *dynamic regret*, which is the difference between the total reward of the best sequence of policies and the total reward of the algorithm. Due to the ubiquity of non-stationary data, there is an increasing trend of designing online algorithms with strong dynamic regret guarantee. We provide a more detailed review of related work in Section 2. In short, while obtaining dynamic regret is relatively well-studied in the full-information setting, for the more challenging bandit feedback, most existing works only focus on the simplest multi-armed bandit problem. More importantly, a sharp contrast between these two regimes is that except for the recent work of (Auer et al., 2018) for a two-armed bandit problem, none of the others achieves optimal dynamic regret *without the knowledge of the non-stationarity of the data* in the

bandit setting, indicating the extra challenge of being adaptive to non-stationary data with partial information.

In this work, we make a significant step in this direction. Specifically we consider the general contextual bandit setting (Auer et al., 2002; Langford and Zhang, 2008) which subsumes many other bandit problems. For an environment with T rounds where at each time t the data is generated from some distribution $\mathcal{D}_t$, denote by $S = 1 + \sum_{t=2}^{T} \mathbf{1}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\}$ the number of switches (plus one) and by $\Delta = \sum_{t=2}^{T} \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}}$ the total variation of these distributions (see Section 3 for more formal definition of the setting) . Our main contribution is to propose an algorithm called ADA-ILTCB$^+$ with the following guarantee:

Main Result ADA-ILTCB$^+$ *achieves the optimal dynamic regret bound* $\mathcal{O}\left(\min\left\{\sqrt{ST}, \Delta^{\frac{1}{3}}T^{\frac{2}{3}}\right\}\right)$ *without knowing* S *or* Δ*. Moreover,* ADA-ILTCB$^+$ *is oracle-efficient.*

Here the dependence on all other parameters are omitted (see Theorem 2 for the complete version) and the optimality of the dependence on S, Δ and T are well-known (Garivier and Moulines, 2011; Besbes et al., 2014). Oracle-efficiency refers to efficiency assuming access to an ERM oracle, a common assumption made in most prior works for efficient contextual bandit (formally defined in Section 3).

Our result is by far the best and most general dynamic regret bound for bandit problems. Recent work by Luo et al. (2018) studies the exact same setting and achieves the same optimal bound only if S and Δ are known; otherwise their algorithms only achieve suboptimal bounds such as $\mathcal{O}(\min\{S^{\frac{1}{4}}T^{\frac{3}{4}}, \Delta^{\frac{1}{5}}T^{\frac{4}{5}}\})$. On the other hand, Auer et al. (2018) propose the first bandit algorithm with expected regret $\mathcal{O}(\sqrt{ST})$ without knowing S, but only for the simplest setting: the two-armed bandit problem without contexts. In contrast, our algorithm works for the general multi-armed bandit problem with contextual information, enjoys a meaningful bound as long as Δ is small (even when S is $\mathcal{O}(T)$), works with high probability, and importantly is oracle-efficient as well.

Our key technique is inspired by (Auer et al., 2018). The high level idea of their algorithm is to occasionally enter some pure exploration phase in order to detect non-stationarity, and crucially the durations of these exploration phases are *multi-scale* and determined in some randomized way. The reason behind this is that smaller non-stationarity requires more time to discover and vice versa. We extend this multi-scale idea to the contextual bandit setting. However, the extension is highly non-trivial and requires the following two new elements:

1. First, we find that pure exploration over arms (used by Auer et al. (2018); Luo et al. (2018)) is not the optimal way to detect non-stationarity in contextual bandit. Instead, we propose to let the algorithm occasionally enter *replay phases*, meaning that the algorithm acts according to some policy distribution used earlier by the algorithm itself. The duration of a replay phase and which previous policy distribution to replay are both determined in some randomized way similar to (Auer et al., 2018). This can be seen as an interpolation between using the current policy distribution and using pure exploration, and as shown by our analysis achieves a better trade off between exploitation and exploration in non-stationary environments.

2. Second, the algorithm of (Auer et al., 2018) is an "arm-elimination" approach, which eliminates arms as long as their sub-optimality is identified. Direct extension to contextual bandit leads to an inefficient approach similar to POLICYELIMINATION by Dudík et al. (2011). Instead, our algorithm is based on the *soft elimination scheme* of (Agarwal et al., 2014) and can

be efficiently implemented with an ERM oracle. Combining this soft elimination scheme and the replay idea in a proper way is another key novelty of our work.

We review related work in Section 2 and introduce all necessary preliminaries in Section 3. Our algorithm is presented in Section 4. The rest of the paper is dedicated to the relatively involved analysis of our algorithm.

2. Related Work

Different forms of dynamic regret bound. Bounding dynamic regret in terms of the number of switches S is traditionally referred to as switching regret or tracking regret, and has been studied under various settings. Note, however, that in some works S refers to the number of switches of data distributions just as our definition (e.g. (Garivier and Moulines, 2011; Wei et al., 2016; Liu et al., 2018; Luo et al., 2018)), while in others S refers to the more general notion of number of switches in the competitor sequence (e.g. (Herbster and Warmuth, 1998; Bousquet and Warmuth, 2002; Auer et al., 2002; Hazan and Seshadhri, 2009)).

Bounding dynamic regret in terms of the variation of loss functions or data distributions is also widely studied (e.g. (Besbes et al., 2014, 2015; Luo et al., 2018)), and there are in fact several other forms of dynamic regret bounds studied in the literature (e.g. (Zinkevich, 2003; Slivkins and Upfal, 2008; Jadbabaie et al., 2015; Wei et al., 2016; Yang et al., 2016; Zhang et al., 2017)).

Adaptivity to non-stationarity. Achieving optimal dynamic regret bounds without any prior knowledge of the non-stationarity is the main focus of this work. This has been achieved for most full-information problems (Luo and Schapire, 2015; Jun et al., 2017; Zhang et al., 2018), but is much more challenging in the bandit setting. Several recent attempts only achieve suboptimal bounds (Karnin and Anava, 2016; Luo et al., 2018; Cheung et al., 2019). It was not clear whether optimal bounds were achievable in this case, until the recent work of Auer et al. (2018) answers this in the affirmative for the two-armed bandit problem. As mentioned our results significantly generalize their work.

Contextual bandits. Contextual bandit is a generalization of the multi-armed bandit problem. While direct generalization of the classic multi-armed bandit algorithm already achieves the optimal static regret (Auer et al., 2002), recent research has been focusing on developing practically efficient algorithms with strong regret guarantee due to their applicability to real-world applications. To avoid running time that is linear in the size of the policy set, most existing works make the practical assumption that an ERM oracle is given to solve the corresponding offline problem. Based on this assumption, a series of progress has been made on developing oracle-efficient algorithms with small static regret (Langford and Zhang, 2008; Dudík et al., 2011; Agarwal et al., 2014; Syrgkanis et al., 2016a; Rakhlin and Sridharan, 2016; Syrgkanis et al., 2016b). All these results rely on some stationary assumption of the environment, since it is known that minimizing static regret oracle-efficiently is impossible in an adversarial environment (Hazan and Koren, 2016).

Despite the negative result for static regret with oracle-efficient algorithms, Luo et al. (2018) find that this is no longer true for dynamic regret, and develop oracle-efficient algorithms with optimal dynamic regret when the non-stationarity is known. Their work is most closely related to ours and our algorithm is in essence similar to their ADA-ILTCB algorithm. The key novelty compared to theirs is the replay phases mentioned earlier, which eventually allows the algorithm to adapt to the non-stationarity of the data.

Replay phases. Introducing replay phases is one of our key contributions. The closest idea in the literature is the method of "mixing past posteriors" of (Bousquet and Warmuth, 2002; Adamskiy et al., 2012), which at each time acts according to some weighted combination of all previous distributions. One key difference of our method is that once it enters into a replay phase, it has to continue for a certain amount of time to gather enough information for non-stationarity detection. Another difference is that in (Bousquet and Warmuth, 2002; Adamskiy et al., 2012) the main point of "mixing past posteriors" is to obtain some form of "long-term memory"; otherwise for typical dynamic regret bounds it is enough to just mix with some amount of pure exploration. It is not clear to us whether our replay idea actually equips the algorithm with some kind of "long-term memory" as well, and we leave this as a future direction.

3. Preliminaries

The contextual bandit problem is defined as follows. Let $\mathcal{X}$ be some arbitrary context space and K be the number of actions. A policy $\pi : \mathcal{X} \to [K]$ is a mapping from the context space to the actions.[1] The learner is given a set of policies Π, assumed to be finite for simplicity but with a huge cardinality $|\Pi|$. Before the learning procedure starts, the environment decides T distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ on $\mathcal{X} \times [0,1]^K$, and draws T independent samples from them: $(x_t, r_t) \sim \mathcal{D}_t, \forall t \in [T]$.[2] The learning procedure then proceeds as follows: for each time $t = 1, \ldots, T$, the learner first receives the context x_t, and then based on this context picks an action $a_t \in [K]$. Afterwards the learner receives the reward feedback $r_t(a_t)$ for the selected action but not others. The instantaneous regret against a policy π at time t is $r_t(\pi(x_t)) - r_t(a_t)$. The classic goal of contextual bandit algorithms is to minimize $\max_{\pi \in \Pi} \sum_{t=1}^{T} r_t(\pi(x_t)) - r_t(a_t)$, that is, the cumulative regret against the best fixed policy, and the optimal bound is known to be $\mathcal{O}(\sqrt{KT \ln |\Pi|})$ in expectation (Auer et al., 2002).

The classic regret is not a good performance measure for non-stationary environments where no single policy can perform well all the time. Instead, we consider dynamic regret that compares the reward of the algorithm to the reward of the best policy *at each time*. Specifically, denote the expected reward of policy π at time t as $\mathcal{R}_t(\pi) \triangleq \mathbb{E}_{(x,r)\sim\mathcal{D}_t}[r(\pi(x))]$, and the optimal policy at time t as $\pi_t^* \triangleq \operatorname{argmax}_{\pi \in \Pi} \mathcal{R}_t(\pi)$. The dynamic regret is then defined as $\sum_{t=1}^{T} r_t(\pi_t^*(x_t)) - r_t(a_t)$.

It is well-known that in general it is impossible to achieve sub-linear dynamic regret. Instead, typical dynamic regret bounds are expressed in terms of some quantities that characterize the non-stationarity of the data distributions, and are meaningful as long as these quantities are sublinear in T. Two such quantities considered in this work are: the number of distribution hard switches (plus one) $S \triangleq 1 + \sum_{t=2}^{T} \mathbf{1}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\}$ and the total variation of distributions $\Delta \triangleq \sum_{t=2}^{T} \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}} = \sum_{t=2}^{T} \int_{[0,1]^K} \int_{\mathcal{X}} \left|\mathcal{D}_t(x,r) - \mathcal{D}_{t-1}(x,r)\right| dx dr$.

More notation. For any integer $1 \leq s \leq s' \leq T$, we denote by $[s, s']$ the time interval $\{s, s+1, \ldots, s'\}$. For an interval $\mathcal{I} = [s, s']$, we define the number of switches and the variation on this interval respectively as $S_{\mathcal{I}} \triangleq 1 + \sum_{\tau=s+1}^{s'} \mathbf{1}\{\mathcal{D}_\tau \neq \mathcal{D}_{\tau-1}\}$ and $\Delta_{\mathcal{I}} \triangleq \sum_{\tau=s+1}^{s'} \|\mathcal{D}_\tau - \mathcal{D}_{\tau-1}\|_{\text{TV}}$.

As in most algorithms, at each time t we sample an action a_t according to some distribution p_t, calculated based on the history before time t. After receiving the reward feedback $r_t(a_t)$, we

1. Throughout the paper we use the notation $[n]$ to denote the set $\{1, \ldots, n\}$ for some integer n.
2. Technically $\mathcal{D}_1, \ldots, \mathcal{D}_T$ are density functions assumed to be absolutely continuous.

construct the usual importance-weighted estimator $\widehat{r}_t$, which is defined as $\widehat{r}_t(a) \triangleq \frac{r_t(a)}{p_t(a)}\mathbf{1}\{a_t = a\}, \forall a \in [K]$ and is clearly unbiased with mean r_t.

For any interval $\mathcal{I} \subseteq [T]$, we define the average reward of a policy π over this interval as $\mathcal{R}_\mathcal{I}(\pi) \triangleq \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \mathcal{R}_t(\pi)$ and similarly its empirical average reward as $\widehat{\mathcal{R}}_\mathcal{I}(\pi) \triangleq \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \widehat{r}_t(\pi(x_t))$. The optimal policy in interval $\mathcal{I}$ is defined as $\pi^*_\mathcal{I} \triangleq \operatorname{argmax}_{\pi\in\Pi} \mathcal{R}_\mathcal{I}(\pi)$ while the empirically best policy is $\widehat{\pi}_\mathcal{I} \triangleq \operatorname{argmax}_{\pi\in\Pi} \widehat{\mathcal{R}}_\mathcal{I}(\pi)$. Furthermore, the expected and empirical interval (static) regret of a policy π for an interval $\mathcal{I}$ are respectively defined as $\mathrm{Reg}_\mathcal{I}(\pi) \triangleq \mathcal{R}_\mathcal{I}(\pi^*_\mathcal{I}) - \mathcal{R}_\mathcal{I}(\pi)$ and $\widehat{\mathrm{Reg}}_\mathcal{I}(\pi) \triangleq \widehat{\mathcal{R}}_\mathcal{I}(\widehat{\pi}_\mathcal{I}) - \widehat{\mathcal{R}}_\mathcal{I}(\pi)$. When $\mathcal{I} = [t, t]$, we simply use t to replace $\mathcal{I}$ as the subscript. For example, $\mathrm{Reg}_t(\pi)$ represents $\mathrm{Reg}_{[t,t]}(\pi)$.

For a context x and a distribution over the policies $Q \in \Delta^\Pi \triangleq \{Q \in \mathbb{R}_+^{|\Pi|} : \sum_{\pi\in\Pi} Q(\pi) = 1\}$, the projected distribution over the actions is denoted by $Q(\cdot|x)$ such that $Q(a|x) = \sum_{\pi:\pi(x)=a} Q(\pi)$ for all $a \in [K]$. The smoothed projected distribution with a minimum probability $\nu \in (0, 1/K]$ is defined as $Q^\nu(\cdot|x) = \nu\mathbf{1} + (1 - K\nu)Q(\cdot|x)$ where $\mathbf{1}$ is the all-one vector. Similarly to (Agarwal et al., 2014), our algorithm keeps track of a bound on the variance of the reward estimates. To this end, define for a policy π, an interval $\mathcal{I}$, a distribution Q, and a minimum probability ν, the empirical and expected variance as

$$\widehat{V}_\mathcal{I}(Q, \nu, \pi) \triangleq \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\left[\frac{1}{Q^\nu(\pi(x_t)|x_t)}\right], \quad V_\mathcal{I}(Q, \nu, \pi) \triangleq \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\mathbb{E}_{x\sim\mathcal{D}_t^\mathcal{X}}\left[\frac{1}{Q^\nu(\pi(x)|x)}\right],$$

where $\mathcal{D}_t^\mathcal{X}$ is the marginal distribution of $\mathcal{D}_t$ over the context space $\mathcal{X}$. Again, $\widehat{V}_t$ and V_t are shorthands for $\widehat{V}_{[t,t]}$ and $V_{[t,t]}$ respectively.

We are interested in efficient algorithms assuming access to an ERM oracle (Agarwal et al., 2014), defined as:

Definition 1 *An ERM oracle is an algorithm which takes any set $\mathcal{T}$ of context-reward pairs $(x, r) \in \mathcal{X} \times \mathbb{R}^K$ as inputs and outputs any policy in $\operatorname{argmax}_{\pi\in\Pi}\sum_{(x,r)\in\mathcal{T}} r(\pi(x))$.*

An algorithm is oracle-efficient if its total running time and the number of oracle calls are both polynomial in T, K and $\ln|\Pi|$, excluding the running time of the oracle itself.

Finally, we use notation $\widetilde{\mathcal{O}}(\cdot)$ to suppress logarithmic dependence on T, K, and $1/\delta$ for some confidence level δ. For notational convenience we also define $\overline{K} = (\log_2 T)K$. A complete notation table can be found in Appendix E.

4. Algorithm

Our algorithm is built upon ILOVETOCONBANDITS of (Agarwal et al., 2014). The main idea of their algorithm is to find a sparse distribution over the policies with both low empirical regret and low empirical variance on the collected data, and then sample actions according to this distribution. Finding such distributions is formalized in Figure 1, Optimization Problem (OP), and Agarwal et al. (2014) show that this can be efficiently implemented using an ERM oracle and importantly the distribution is sparse. Under a stationary environment, it can be shown that the empirical regret concentrates around the expected regret reasonably well and thus the algorithm has low regret.

3. We emphasize "sparse distribution" only to ensure the efficiency of the algorithm. Whether $Q_{(i,j)}$ is sparse or not does not affect the regret bound since we trivially bound the regret for block 0 by its length.

Algorithm 1 ADA-ILTCB$^+$

Input: confidence level $\delta \in (0,1)$, time horizon T, underlying policy class Π.

Definition: $\nu_j = \sqrt{\frac{C_0}{K2^jL}}$, where $C_0 = \ln\left(\frac{8T^3|\Pi|^2}{\delta}\right)$ and $L = \lceil 4KC_0 \rceil$ (block base length).

$\mathcal{B}_{(i,j)} \triangleq [\tau_i, \tau_i + 2^jL - 1]$, where τ_i is the beginning of epoch i, as defined in the algorithm.

1 **Initialize:** $t = 1, i = 1$.
2 $\tau_i \leftarrow t$. ▷ i indexes an epoch
3 **for** $j = 0, 1, 2, \ldots$ **do** ▷ j indexes a block
4 If $j = 0$, define $Q_{(i,j)}$ as an arbitrary sparse distribution over Π;[3] otherwise, let $Q_{(i,j)}$ be the solution of (OP) with inputs $\mathcal{I} = \mathcal{B}_{(i,j-1)}$ and $\nu = \nu_j$.
5 $\mathcal{S} \leftarrow \emptyset$. ▷ $\mathcal{S}$ records replay indices and intervals
6 **while** $t \le \tau_i + 2^jL - 1$ **do**
7 ▷ Step 1. Randomly start a replay phase
8 Sample REP$\sim$ Bernoulli $\left(\frac{1}{L} \times 2^{-j/2} \times \sum_{m=0}^{j-1} 2^{-m/2}\right)$.
9 **if** REP $= 1$ **then**
10 Sample m from $\{0, \ldots, j-1\}$ s.t. $\Pr[m = b] \propto 2^{-b/2}$.
11 $\mathcal{S} \leftarrow \mathcal{S} \cup \{(m, [t, t + 2^mL - 1])\}$. ▷ start a new replay phase
12 ▷ Step 2. Sample an action
13 Let $M_t \triangleq \{m \mid \exists\, \mathcal{I}$ such that $t \in \mathcal{I}$ and $(m, \mathcal{I}) \in \mathcal{S}\}$.
14 **if** M_t *is empty* **then**
15 Play $a_t \sim Q_{(i,j)}^{\nu_j}(\cdot|x_t)$.
16 **else**
17 Sample $m \sim$ Uniform(M_t)
18 Play $a_t \sim Q_{(i,m)}^{\nu_m}(\cdot|x_t)$.
19 ▷ Step 3. Perform non-stationarity tests
20 **for** $(m, [s, s']) \in \mathcal{S}$ **do**
21 **if** $s' = t$ **and** ENDOFREPLAYTEST $(i, j, m, [s, t]) =$ *Fail* **then**
22 $t \leftarrow t + 1$, $i \leftarrow i + 1$ and **goto** Line 2 to start a new epoch.
23 **if** $t = \tau_i + 2^jL - 1$ **and** ENDOFBLOCKTEST $(i, j) =$*Fail* **then**
24 $t \leftarrow t + 1$, $i \leftarrow i + 1$ and **goto** Line 2 to start a new epoch.
25 $t \leftarrow t + 1$.

The ADA-ILTCB algorithm of (Luo et al., 2018) works by equipping ILOVETOCONBANDITS with some non-stationarity tests and restarting once non-stationarity is detected. Our algorithm ADA-ILTCB$^+$ works under a similar framework with similar tests, but importantly enters into replay phases occasionally. The complete pseudocode is included in Algorithm 1 and we describe in detail how it works below.

The algorithm starts a new *epoch* every time it restarts (that is, on execution of Line 22 or 24). We index an epoch by i and denote the first round of epoch i by τ_i. Within an epoch, the algorithm

Optimization Problem (OP)
Input: time interval $\mathcal{I}$, minimum exploration probability ν
Return $Q \in \Delta^{\Pi}$ such that for constant $C = 1.2 \times 10^7$,

$$\sum_{\pi \in \Pi} Q(\pi) \widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) \leq 2CK\nu, \tag{1}$$

$$\forall \pi \in \Pi : \widehat{V}_{\mathcal{I}}(Q, \nu, \pi) \leq 2K + \frac{\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi)}{C\nu}. \tag{2}$$

ENDOFREPLAYTEST$(i, j, m, \mathcal{A})$
Return *Fail* if there exists $\pi \in \Pi$ such that any of the following inequalities holds:

$$\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) - 4\widehat{\mathrm{Reg}}_{\mathcal{B}_{(i,j-1)}}(\pi) \geq D_1 \overline{K} \nu_m, \tag{3}$$

$$\widehat{\mathrm{Reg}}_{\mathcal{B}_{(i,j-1)}}(\pi) - 4\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) \geq D_1 \overline{K} \nu_m, \tag{4}$$

$$\widehat{V}_{\mathcal{A}}(Q_{(i,m)}, \nu_m, \pi) - 41\widehat{V}_{\mathcal{B}_{(i,j-1)}}(Q_{(i,m)}, \nu_m, \pi) \geq D_2 K, \tag{5}$$

where $D_1 = 6400$ and $D_2 = 800$; otherwise return *Pass*.

ENDOFBLOCKTEST(i, j)
Return *Fail* if there exists $k \in \{0, \ldots, j-1\}$ and $\pi \in \Pi$ such that any of the following inequalities holds:

$$\widehat{\mathrm{Reg}}_{\mathcal{B}_{(i,j)}}(\pi) - 4\widehat{\mathrm{Reg}}_{\mathcal{B}_{(i,k)}}(\pi) \geq D_4 \overline{K} \nu_k, \tag{6}$$

$$\widehat{\mathrm{Reg}}_{\mathcal{B}_{(i,k)}}(\pi) - 4\widehat{\mathrm{Reg}}_{\mathcal{B}_{(i,j)}}(\pi) \geq D_4 \overline{K} \nu_k, \tag{7}$$

$$\widehat{V}_{\mathcal{B}_{(i,j)}}(Q_{(i,k+1)}, \nu_{k+1}, \pi) - 41\widehat{V}_{\mathcal{B}_{(i,k)}}(Q_{(i,k+1)}, \nu_{k+1}, \pi) \geq D_5 K, \tag{8}$$

where $D_4 = 6400$ and $D_5 = 800$; otherwise return *Pass*.

Figure 1: Optimization subroutine and non-stationarity tests

works on a *block* schedule. Specifically, in epoch i, we call the interval $[\tau_i, \tau_i + L - 1]$ block 0 and interval $[\tau_i + 2^{j-1}L, \tau_i + 2^j L - 1]$ block j for any $j \geq 1$ (in the case of restart, the block ends earlier), where L is some fixed base length.[4] Each block is associated with an exploration probability ν_j of order $1/\sqrt{K 2^j L}$. At the beginning of each block j (for $j \geq 1$), the algorithm first solves the Optimization Problem (OP) (Figure 1) using exploration probability ν_j and all data

4. The lengths of these blocks are doubling except that block 0 and block 1 have the same length L. This is merely for notational convenience and it is not crucial.

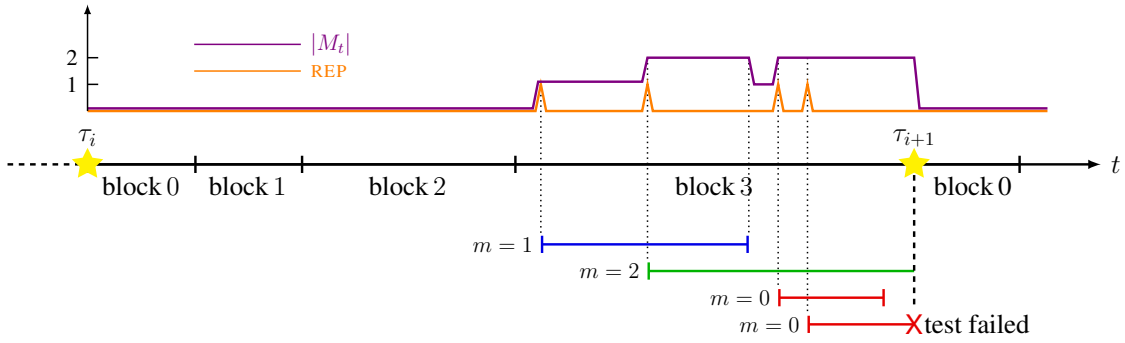

Figure 2: An Illustration of ADA-ILTCB$^+$ (best viewed in color). The two stars represent two restarts and the interval between them is epoch i. The purple curve represents the value of $|M_t|$, that is, how many replay phases (with distinct indices) the algorithm is currently in. The orange curve represents the value of the Bernoulli variable REP, so that each spike represents the start of a new replay interval. The four segments below the t-axis indicates four replay intervals started within block 3, with their index value m on the left. Segments with the same index share the same color, and segment with larger index is longer. At time $\tau_{i+1} - 1$, the bottom replay interval finishes and the algorithm performs a ENDOFREPLAYTEST. The test fails and the algorithm restarts a new epoch. Note that the green replay interval (with $m = 2$) is also discontinued due to the restart.

collected since the beginning of the current epoch, that is, data from $\mathcal{B}_{(i,j-1)} \triangleq [\tau_i, \tau_i + 2^{j-1}L - 1]$. The solution is denoted by $Q_{(i,j)}$, which is a sparse distribution over policies.

Afterwards, for most of the time of the current block, the algorithm simply plays according to $Q^{\nu_j}_{(i,j)}(\cdot|x_t)$, just like ILOVETOCONBANDITS. The difference is that at each time, with probability $\frac{1}{L}2^{-j/2}2^{-m/2}$ the algorithm enters into a *replay phase* of index $m \in \{0, 1, \ldots, j-1\}$ which lasts for $2^m L$ rounds. This is implemented in Line 8-11, where we first sample a Bernoulli variable REP to decide whether or not to enter into a replay phase, and if so then randomly select a replay index m to ensure the aforementioned probability. The set $\mathcal{S}$ is used to record all pairs of replay index and replay interval. Similar to (Auer et al., 2018), the reason of using different lengths is to allow the algorithm to detect different level of non-stationarity: a longer replay interval with a larger index is used to detect smaller non-stationarity.

Note that at each time t, the algorithm could potentially be in multiple replay phases simultaneously. Let M_t be the set of indices of all the *ongoing* replay intervals (defined in Line 13). If M_t is empty, the algorithm is not in any replay phase and simply samples an action according to $Q^{\nu_j}_{(i,j)}(\cdot|x_t)$ as mentioned. On the other hand, if M_t is not empty, the algorithm uniformly at random picks an index m from M_t, and then replays the distribution learned at the beginning of block m, that is, samples an action according to $Q^{\nu_m}_{(i,m)}(\cdot|x_t)$. Recall that our reward estimators $\widehat{r}_t$'s are defined in terms of a distribution p_t over actions, and it is clear that for our algorithm $p_t(\cdot) = \mathbf{1}\{|M_t| = 0\}Q^{\nu_j}_{(i,j)}(\cdot|x_t) + \mathbf{1}\{|M_t| \neq 0\}\frac{1}{|M_t|}\sum_{m \in M_t} Q^{\nu_m}_{(i,m)}(\cdot|x_t)$.

Finally, at the end of every replay interval, the algorithm calls the subroutine ENDOFREPLAYTEST to check whether the data collected in the replay interval and that collected prior to

the current block (that is, $\mathcal{B}_{(i,j-1)}$) are consistent (Line 21). Also, at the end of every block j, the algorithm calls another subroutine ENDOFBLOCKTEST to check the consistency between data up to block j and data up to block k for all $k \in \{0, 1, \ldots, j-1\}$ (Line 23). Both tests are in similar spirit to those of (Luo et al., 2018), and check the difference of empirical regret or empirical variance of each policy over different sets of data (see Figure 1). The difference between the two tests is that they capture non-stationarity at different time steps. If either of the tests indicates that there is a significant distribution change, the algorithm restarts from scratch and enters into the next epoch. Also note that if ENDOFBLOCKTEST passes and the algorithm enters into a new block, all unfinished replay intervals will discontinue ($\mathcal{S}$ is reset to be empty in Line 5).

We provide an illustration of our algorithm in Figure 2.

Oracle-efficiency. Our algorithm can be implemented efficiently with an ERM oracle. Agarwal et al. (2014) show that the Optimization Problem (OP) with input ν can be solved using $\widetilde{\mathcal{O}}(1/\nu)$ oracle calls with a solution that is $\widetilde{\mathcal{O}}(1/\nu)$-sparse. In our case, $\widetilde{\mathcal{O}}(1/\nu)$ is at most $\widetilde{\mathcal{O}}(\sqrt{KT})$. The two tests can also be implemented efficiently by the exact same arguments of (Luo et al., 2018). For example, in ENDOFREPLAYTEST, to check if there exists a $\pi \in \Pi$ satisfying Eq. (3), we can first use two oracle calls to precompute $\max_{\pi' \in \Pi} \widehat{\mathcal{R}}_{\mathcal{A}}(\pi')$ and $\max_{\pi' \in \Pi} \widehat{\mathcal{R}}_{\mathcal{B}_{(i,j-1)}}(\pi')$, and collect $\mathcal{T} = \left\{\left(x_t, \frac{-\widehat{r}_t}{|\mathcal{A}|}\right)\right\}_{t \in \mathcal{A}} \cup \left\{\left(x_t, \frac{4\widehat{r}_t}{|\mathcal{B}_{(i,j-1)}|}\right)\right\}_{t \in \mathcal{B}_{(i,j-1)}}$. Then we again use an oracle call to find $\max_{\pi \in \Pi} \sum_{(x,r) \in \mathcal{T}} r(\pi(x))$ and add this value to $\max_{\pi' \in \Pi} \widehat{\mathcal{R}}_{\mathcal{A}}(\pi') - 4 \max_{\pi' \in \Pi} \widehat{\mathcal{R}}_{\mathcal{B}_{(i,j-1)}}(\pi')$, which is equal to taking the max over $\pi \in \Pi$ of the left hand side of Eq. (3). It remains to compare this value with the right hand side of Eq. (3).

5. Main Theorem and Proof Outline

The dynamic regret guarantee of ADA-ILTCB$^+$ is summarized below:

Theorem 2 (Main Theorem) ADA-ILTCB$^+$ *guarantees with high probability,*

$$\sum_{t=1}^{T} r_t(\pi_t^*(x_t)) - r_t(a_t) = \widetilde{\mathcal{O}}\left(\min\left\{\sqrt{K(\ln|\Pi|)ST}, \sqrt{K(\ln|\Pi|)T} + (K\ln|\Pi|)^{\frac{1}{3}}\Delta^{\frac{1}{3}}T^{\frac{2}{3}}\right\}\right).$$

Proof roadmap. The rest of the paper proves our main theorem, following these steps: in Section 5.1, we provide a key lemma that bounds the dynamic regret for any interval within a block (in terms of some algorithm-dependent quantities). In Section 5.2, with the help of the key lemma we bound the dynamic regret for a block. In Section 5.3 we bound the number of epochs/restarts, and sum up the regret over all blocks in all epochs to get the final bound. Since the analysis in Sections 5.1 and 5.2 is all about a fixed epoch i, for notation simplicity, we simply write $\mathcal{B}_{(i,j)}$ and $Q_{(i,j)}$ as $\mathcal{B}_j$ and Q_j in these two sections.

5.1. A main Lemma and regret decomposition

To bound the dynamic regret over any interval, we define the concept of *excess regret*:

Definition 3 *For an interval $\mathcal{I}$ that lies in $[\tau_i + 2^{j-1}L, \tau_i + 2^j L - 1]$ for some $j > 0$, we define its excess regret as*

$$\varepsilon_{\mathcal{I}} \triangleq \max_{\pi \in \Pi} \mathrm{Reg}_{\mathcal{I}}(\pi) - 8\widehat{\mathrm{Reg}}_{\mathcal{B}_{(i,j-1)}}(\pi),$$

and its excess regret threshold as $\alpha_{\mathcal{I}} = \sqrt{\frac{2KC_0}{|\mathcal{I}|}} \log_2 T$.

In words, excess regret of $\mathcal{I}$ is the maximum discrepancy between a policy's expected static regret on $\mathcal{I}$ and (8 times) its empirical static regret on the first j blocks. Large excess regret thus indicates non-stationarity. We now use the following main lemma to decompose the dynamic regret on $\mathcal{I}$ based on whether the excess regret reaches the excess regret threshold.

Lemma 4 (Main Lemma) *With probability* $1-\delta$, ADA-ILTCB$^+$ *guarantees for all* $j > 0$ *and any interval* $\mathcal{I}$ *that lies in block* j,

$$\sum_{t\in\mathcal{I}} (r_t(\pi_t^*(x_t)) - r_t(a_t)) = \mathcal{O}\left(\left(\sum_{t\in\mathcal{I}} \sum_{m\in M_t\cup\{j\}} \overline{K}\nu_m\right) + |\mathcal{I}|\alpha_{\mathcal{I}} + |\mathcal{I}|\Delta_{\mathcal{I}} + |\mathcal{I}|\varepsilon_{\mathcal{I}}\mathbf{1}\{\varepsilon_{\mathcal{I}} > D_3\alpha_{\mathcal{I}}\}\right)$$

where $D_3 = 4.1 \times 10^6$.

Proof By Azuma's inequality and a union bound over all T^2 possible intervals, we have that with probability $1-\delta$, for any interval $\mathcal{I}$,

$$\sum_{t\in\mathcal{I}} (r_t(\pi_t^*(x_t)) - r_t(a_t)) \le \sum_{t\in\mathcal{I}} \mathbb{E}_t[r_t(\pi_t^*(x_t)) - r_t(a_t)] + \mathcal{O}\left(\sqrt{|\mathcal{I}|\log(T^2/\delta)}\right), \tag{9}$$

where $\mathbb{E}_t$ is the conditional expectation given everything up to Step 1 of the algorithm of round t. It remains to bound each $\mathbb{E}_t[r_t(\pi_t^*(x_t)) - r_t(a_t)]$. Depending on the case of replay or non-replay, this term can be written as

$$\mathbb{E}_t\left[r_t(\pi_t^*(x_t)) - r_t(a_t)\right] = \begin{cases} \sum_{a\in[K]} \sum_{m\in M_t} \frac{Q_m^{\nu_m}(a|x_t)}{|M_t|} \mathbb{E}_t\left[r_t(\pi_t^*(x_t)) - r_t(a)\right], & \text{if } M_t \neq \emptyset, \\ \sum_{a\in[K]} Q_j^{\nu_j}(a|x_t)\mathbb{E}_t\left[r_t(\pi_t^*(x_t)) - r_t(a)\right], & \text{if } M_t = \emptyset. \end{cases}$$

Now observe that for any Q and ν, by definition of Q^ν we have

$$\sum_{a\in[K]} Q^\nu(a|x_t)\mathbb{E}_t\left[r_t(\pi_t^*(x_t)) - r_t(a)\right] \le K\nu + \sum_{\pi\in\Pi} Q(\pi)\text{Reg}_t(\pi).$$

So we continue to bound $\mathbb{E}_t\left[r_t(\pi_t^*(x_t)) - r_t(a_t)\right]$ by

$$\sum_{m\in M_t\cup\{j\}} K\nu_m + \begin{cases} \frac{1}{|M_t|}\sum_{m\in M_t}\sum_{\pi\in\Pi} Q_m(\pi)\text{Reg}_t(\pi), & \text{if } M_t \neq \emptyset, \\ \sum_{\pi\in\Pi} Q_j(\pi)\text{Reg}_t(\pi), & \text{if } M_t = \emptyset. \end{cases} \tag{10}$$

Next note that for any $t \in \mathcal{I}$ and $m \in \{1,\dots,j\}$, we have

$$\begin{aligned} \sum_{\pi\in\Pi} Q_m(\pi)\text{Reg}_t(\pi) &\le \sum_{\pi\in\Pi} Q_m(\pi)\text{Reg}_{\mathcal{I}}(\pi) + \mathcal{O}(\Delta_{\mathcal{I}}) && \text{(Lemma 8)} \\ &= \sum_{\pi\in\Pi} 8Q_m(\pi)\widehat{\text{Reg}}_{\mathcal{B}_{j-1}}(\pi) + \mathcal{O}(\Delta_{\mathcal{I}}) + \varepsilon_{\mathcal{I}} && \text{(definition of } \varepsilon_{\mathcal{I}}) \end{aligned}$$

$$\leq \sum_{\pi\in\Pi} 8Q_m(\pi)\left(4\widehat{\mathrm{Reg}}_{\mathcal{B}_{m-1}}(\pi) + D_4\overline{K}\nu_m\right) + \mathcal{O}(\Delta_{\mathcal{I}}) + \varepsilon_{\mathcal{I}}$$

(Eq. (6) does not hold)

$$\leq \mathcal{O}(\overline{K}\nu_m + \Delta_{\mathcal{I}}) + \varepsilon_{\mathcal{I}} \qquad \text{(Eq. (1))}$$

$$\leq \mathcal{O}(\overline{K}\nu_m + \alpha_{\mathcal{I}} + \Delta_{\mathcal{I}}) + \varepsilon_{\mathcal{I}}\mathbf{1}\left\{\varepsilon_{\mathcal{I}} > D_3\alpha_{\mathcal{I}}\right\}.$$

In fact, the above holds for $m = 0$ too since the left hand side is at most $1 \leq 4\overline{K}\nu_0$. Combining this inequality with Eq. (10) and (9), and noting that the term $\sqrt{|\mathcal{I}|\log(T^2/\delta)}$ is of order $\mathcal{O}\left(|\mathcal{I}|\alpha_{\mathcal{I}}\right)$ finish the proof. ■

5.2. Dynamic regret for a block

In this section, we bound the dynamic regret of some block $j > 0$ within epoch i. This block can be formally written as

$$\mathcal{J} \triangleq [\tau_i, \tau_{i+1} - 1] \cap \left[\tau_i + 2^{j-1}L, \tau_i + 2^j L - 1\right]. \tag{11}$$

The idea is to divide $\mathcal{J}$ into several intervals, apply Lemma 4 to each of them, and finally sum up the regret. Importantly, we need to divide $\mathcal{J}$ in a careful way according to the following lemma, so that the variation on each interval is bounded by its excess regret threshold, while at the same time the number of intervals is not too large. Note that this division only happens in the analysis.

Lemma 5 *There is a way to partition any interval $\mathcal{J}$ into $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_\Gamma$, such that $\Delta_{\mathcal{I}_k} \leq \alpha_{\mathcal{I}_k}, \forall k \in [\Gamma]$, and $\Gamma = \mathcal{O}(\min\{S_{\mathcal{J}}, (KC_0)^{-\frac{1}{3}}\Delta_{\mathcal{J}}^{\frac{2}{3}}|\mathcal{J}|^{\frac{1}{3}} + 1\})$.*

For the first $\Gamma - 1$ intervals of this partition, we apply Lemma 4 to each of them. Note that the term $|\mathcal{I}_k|\Delta_{\mathcal{I}_k}$ in Lemma 4 can be absorbed by the term $|\mathcal{I}_k|\alpha_{\mathcal{I}_k}$ by our partition property. Summing up the bounds from Lemma 4, we get the following dynamic regret bound for these $\Gamma - 1$ intervals:

$$\sum_{k=1}^{\Gamma-1}\mathcal{O}\left(\left(\sum_{t\in\mathcal{I}_k}\sum_{m\in M_t\cup\{j\}}\overline{K}\nu_m\right) + |\mathcal{I}_k|\alpha_{\mathcal{I}_k} + |\mathcal{I}_k|\Delta_{\mathcal{I}_k} + |\mathcal{I}_k|\varepsilon_{\mathcal{I}_k}\mathbf{1}\{\varepsilon_{\mathcal{I}_k} > D_3\alpha_{\mathcal{I}_k}\}\right)$$

$$\leq \underbrace{\sum_{k=1}^{\Gamma-1}\sum_{t\in\mathcal{I}_k}\sum_{m\in M_t\cup\{j\}}\mathcal{O}(\overline{K}\nu_m)}_{\textsc{Term}_1} + \underbrace{\sum_{k=1}^{\Gamma-1}\mathcal{O}(|\mathcal{I}_k|\alpha_{\mathcal{I}_k})}_{\textsc{Term}_2} + \underbrace{\sum_{k=1}^{\Gamma-1}\mathcal{O}(|\mathcal{I}_k|\varepsilon_{\mathcal{I}_k}\mathbf{1}\{\varepsilon_{\mathcal{I}_k} > D_3\alpha_{\mathcal{I}_k}\})}_{\textsc{Term}_3}. \tag{12}$$

For the last interval in the block, it is possible that it was interrupted by a restart, which makes the analysis trickier and we defer the details to Appendix C. Further bounding $\textsc{Term}_1$ and $\textsc{Term}_2$ is relatively straightforward by the definition of ν_m and $\alpha_{\mathcal{I}_k}$ and also the construction of M_t (see Appendix C). For $\textsc{Term}_3$, the idea is that this term is nonzero only when $\varepsilon_{\mathcal{I}_k}$ is large, which implies that the distribution in $\mathcal{I}_k$ is quite different from that in $\mathcal{B}_{j-1}$. In this case we will show that as long as the algorithm starts a replay phase with some "correct" index within $\mathcal{I}_k$, it will detect the non-stationarity with high probability and restart the algorithm. Thus we only need to bound the regret accumulated before this "correct" replay phase appears. We provide the complete proof in Appendix C.1, which is the most important part of the analysis. Combining the bounds for these three terms, we eventually arrive at the following lemma:

Lemma 6 *With probability $1-\delta$, the following holds for any block $\mathcal{J}$ with block index $j>0$:*

$$\sum_{t\in\mathcal{J}}(r_t(\pi_t^*)-r_t(a_t))=\widetilde{\mathcal{O}}\left(\min\left\{\sqrt{KC_0S_{\mathcal{J}}2^jL},\sqrt{KC_02^jL}+(KC_0)^{\frac{1}{3}}\Delta_{\mathcal{J}}^{\frac{1}{3}}(2^jL)^{\frac{2}{3}}\right\}\right).$$

Note that 2^jL is the length of block $\mathcal{J}$ unless there is a restart triggered within this block, in which case the length is smaller.

5.3. Combining regret over blocks and epochs

We finally sum up the dynamic regret over blocks and epochs. To this end, we reintroduce the subscripts i,j in our notations, and write epoch i as $\mathcal{E}_i=[\tau_i,\tau_{i+1}-1]$ and block j (for $j>0$) in epoch i as $\mathcal{J}_{ij}=[\tau_i+2^{j-1}L,\tau_i+2^jL-1]\cap\mathcal{E}_i$.

Dynamic regret for an epoch. The last block index in epoch i is $\max\{0,\lceil\log_2(|\mathcal{E}_i|/L)\rceil\}$, which we denote by j^*. Using Lemma 6, we combine the regret over all blocks in epoch i and upper bound the regret of epoch i simultaneously by (using the bound in terms of number of switches)

$$\widetilde{\mathcal{O}}\left(L+\sum_{j=1}^{j^*}\sqrt{KC_0S_{\mathcal{J}_{ij}}2^jL}\right)=\widetilde{\mathcal{O}}\left(KC_0+\sqrt{KC_0\sum_{j=1}^{j^*}S_{\mathcal{J}_{ij}}\sum_{j=1}^{j^*}2^jL}\right)\quad\text{(Cauchy-Schwarz)}$$
$$=\widetilde{\mathcal{O}}\left(KC_0+\sqrt{KC_0(S_{\mathcal{E}_i}+j^*)|\mathcal{E}_i|}\right)=\widetilde{\mathcal{O}}\left(\sqrt{KC_0S_{\mathcal{E}_i}|\mathcal{E}_i|}\right)$$

and similarly by (using the bound in terms of variation and Hölder inequality)

$$\widetilde{\mathcal{O}}\left(L+\sum_{j=1}^{j^*}\sqrt{KC_02^jL}+\sum_{j=1}^{j^*}(KC_0)^{\frac{1}{3}}\Delta_{\mathcal{J}}^{\frac{1}{3}}(2^jL)^{\frac{2}{3}}\right)=\widetilde{\mathcal{O}}\left(\sqrt{KC_0|\mathcal{E}_i|}+(KC_0)^{\frac{1}{3}}\Delta_{\mathcal{E}_i}^{\frac{1}{3}}|\mathcal{E}_i|^{\frac{2}{3}}\right).$$

Combining regret over epochs. For the last step of combining all epochs, we make use the following lemma which bounds the number of epochs (see Appendix D for the proof).

Lemma 7 *Denote the total number of epochs by E. With probability at least $1-\delta/2$, we have $E\le\min\{S,(KC_0)^{-\frac{1}{3}}\Delta^{\frac{2}{3}}T^{\frac{1}{3}}+1\}$.*

Therefore, summing up the previous bounds over all epochs, we arrive at the final dynamic regret bound, which is the minimum of the following two:

$$\widetilde{\mathcal{O}}\left(\sum_{i=1}^{E}\sqrt{KC_0S_{\mathcal{E}_i}|\mathcal{E}_i|}\right)\le\widetilde{\mathcal{O}}\left(\sqrt{KC_0\left(\sum_{i=1}^{E}S_{\mathcal{E}_i}\right)\left(\sum_{i=1}^{E}|\mathcal{E}_i|\right)}\right)$$
$$=\widetilde{\mathcal{O}}\left(\sqrt{KC_0(S+E)T}\right)=\widetilde{\mathcal{O}}\left(\sqrt{KC_0ST}\right),$$

and by

$$\widetilde{\mathcal{O}}\left(\sum_{i=1}^{E}\left(\sqrt{KC_0|\mathcal{E}_i|}+(KC_0)^{\frac{1}{3}}\Delta_{\mathcal{E}_i}^{\frac{1}{3}}|\mathcal{E}_i|^{\frac{2}{3}}\right)\right)\le\widetilde{\mathcal{O}}\left(\sqrt{KC_0ET}+(KC_0)^{\frac{1}{3}}\left(\sum_{i=1}^{E}\Delta_{\mathcal{E}_i}\right)^{\frac{1}{3}}T^{\frac{2}{3}}\right)$$
$$=\widetilde{\mathcal{O}}\left(\sqrt{KC_0T}+(KC_0)^{\frac{1}{3}}\Delta^{\frac{1}{3}}T^{\frac{2}{3}}\right).$$

This proves the bound stated in the main theorem.

Acknowledgments

The authors would like to thank Peter Auer for the discussion about the possibility of getting optimal bounds for our problem, and to thank Peter Auer, Pratik Gajane, Ronald Ortner for kindly sharing their manuscript of (Auer et al., 2018) before it was public. HL and CYW are supported by NSF Grant #1755781.

References

Dmitry Adamskiy, Manfred K Warmuth, and Wouter M Koolen. Putting bayes to sleep. In *Advances in neural information processing systems 25*, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

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

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

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

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

Olivier Bousquet and Manfred K Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3(Nov):363–396, 2002.

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

M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2011.

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

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, 2016.

Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th International Conference on Machine Learning*, pages 393–400, 2009.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, Rebecca Willett, et al. Online learning for changing environments using coin betting. *Electronic Journal of Statistics*, 11(2):5282–5310, 2017.

Zohar S Karnin and Oren Anava. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems 29*, 2016.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 21*, 2008.

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Haipeng Luo and Robert E. Schapire. Achieving All with No Parameters: AdaNormalHedge. In *28th Annual Conference on Learning Theory (COLT)*, 2015.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *31st Annual Conference on Learning Theory (COLT)*, 2018.

Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *21st Annual Conference on Learning Theory (COLT)*, pages 343–354, 2008.

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert E Schapire. Efficient algorithms for adversarial contextual learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016a.

Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert E Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems 29*, 2016b.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in Neural Information Processing Systems 29*, 2016.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 449–457, 2016.

Lijun Zhang, Tianbao Yang, Jinfeng Yi, Jing Rong, and Zhi-Hua Zhou. Improved dynamic regret for non-degenerate functions. In *Advances in Neural Information Processing Systems 30*, 2017.

Lijun Zhang, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

Appendix A. Useful Lemmas

In this section we prove two small lemmas that are useful for our analysis.

A.1. Discrepancy between intervals

The following results allow us to relate regret and variance measured on one interval to those measured on another, with the price in terms of the distribution variation.

Lemma 8 *For any interval $\mathcal{I}$, its sub-intervals $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathcal{I}$, and any $\pi \in \Pi$, we have*

$$\left|\mathrm{Reg}_{\mathcal{I}_1}(\pi) - \mathrm{Reg}_{\mathcal{I}_2}(\pi)\right| \leq 2\Delta_{\mathcal{I}}.$$

Proof Let $\pi^\star_{\mathcal{I}_1} = \mathrm{argmax}_{\pi\in\Pi} \mathcal{R}_{\mathcal{I}_1}(\pi)$ and $\pi^\star_{\mathcal{I}_2} = \mathrm{argmax}_{\pi\in\Pi} \mathcal{R}_{\mathcal{I}_2}(\pi)$. Then

$$\mathrm{Reg}_{\mathcal{I}_1}(\pi) - \mathrm{Reg}_{\mathcal{I}_2}(\pi) = \mathcal{R}_{\mathcal{I}_1}(\pi^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_1}(\pi) - \mathcal{R}_{\mathcal{I}_2}(\pi^\star_{\mathcal{I}_2}) + \mathcal{R}_{\mathcal{I}_2}(\pi).$$

By definition we have

$$-\Delta_{\mathcal{I}} \leq \mathcal{R}_{\mathcal{I}_2}(\pi) - \mathcal{R}_{\mathcal{I}_1}(\pi) \leq \Delta_{\mathcal{I}},$$

and

$$-\Delta_{\mathcal{I}} \leq \mathcal{R}_{\mathcal{I}_1}(\pi^\star_{\mathcal{I}_2}) - \mathcal{R}_{\mathcal{I}_2}(\pi^*_{\mathcal{I}_2}) \leq \mathcal{R}_{\mathcal{I}_1}(\pi^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_2}(\pi^\star_{\mathcal{I}_2}) \leq \mathcal{R}_{\mathcal{I}_1}(\pi^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_2}(\pi^\star_{\mathcal{I}_1}) \leq \Delta_{\mathcal{I}}.$$

Combining them we get the desired bound. ■

Lemma 9 *For any interval $\mathcal{I}$, its sub-intervals $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathcal{I}$, any $Q \in \Delta^\Pi$, $\mu \in (0, 1/K]$, and $\pi \in \Pi$, we have*

$$\left|V_{\mathcal{I}_1}(Q, \mu, \pi) - V_{\mathcal{I}_2}(Q, \mu, \pi)\right| \leq \frac{\Delta_{\mathcal{I}}}{\mu}.$$

Proof For any $s, t \in \mathcal{I}$ (assuming $s < t$), any P, and $\pi \in \Pi$,

$$\begin{aligned}\left|V_s(Q, \mu, \pi) - V_t(Q, \mu, \pi)\right| &= \left|\mathbb{E}_{\mathcal{D}_s^{\mathcal{X}}}\left[\frac{1}{Q^\mu(\pi(x)|x)}\right] - \mathbb{E}_{\mathcal{D}_t^{\mathcal{X}}}\left[\frac{1}{Q^\mu(\pi(x)|x)}\right]\right| \\ &= \left|\int_{\mathcal{X}} (\mathcal{D}_s^{\mathcal{X}}(x) - \mathcal{D}_t^{\mathcal{X}}(x)) \frac{1}{Q^\mu(\pi(x)|x)} dx\right|\end{aligned}$$

$$\leq \frac{1}{\mu}\int_{\mathcal{X}} \left|\mathcal{D}_s^{\mathcal{X}}(x) - \mathcal{D}_t^{\mathcal{X}}(x)\right| dx$$
$$\leq \frac{1}{\mu}\sum_{\tau=s+1}^{t} \|\mathcal{D}_\tau - \mathcal{D}_{\tau-1}\|_{\text{TV}} \leq \frac{\Delta_{\mathcal{I}}}{\mu}.$$

Therefore,

$$\left|V_{\mathcal{I}_1}(Q,\mu,\pi) - V_{\mathcal{I}_2}(Q,\mu,\pi)\right| \leq \frac{1}{|\mathcal{I}_1|}\frac{1}{|\mathcal{I}_2|}\sum_{s\in\mathcal{I}_1}\sum_{t\in\mathcal{I}_2}\left|V_s(Q,\mu,\pi) - V_t(Q,\mu,\pi)\right| \leq \frac{\Delta_{\mathcal{I}}}{\mu}.$$

■

A.2. Partitioning an interval

We prove Lemma 5 in this section, which states that for any interval $\mathcal{J}$, there exists a way to partition $\mathcal{J}$ into $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_\Gamma$, such that

$$\Delta_{\mathcal{I}_k} \leq \alpha_{\mathcal{I}_k} = \sqrt{\frac{2KC_0}{|\mathcal{I}_k|}}\log_2 T, \quad \forall k \in [\Gamma],$$

and

$$\Gamma = \mathcal{O}\left(\min\left\{S_{\mathcal{J}}, (KC_0)^{-\frac{1}{3}}\Delta_{\mathcal{J}}^{\frac{2}{3}}|\mathcal{J}|^{\frac{1}{3}} + 1\right\}\right).$$

We prove this by giving an explicit greedy "algorithm", but we emphasize that this only happens in the analysis and is never really needed to be executed.

Proof [of Lemma 5] Consider the following partitioning procedure.

Algorithm 2 Partitioning an interval

Input: an interval $\mathcal{J} = [s, e]$.
Initialize: Let $k = 1$, $s_1 = s, t = s$.
while $t \leq e$ **do**
  if $\Delta_{[s_k,t]} \leq \sqrt{\frac{KC_0}{t-s_k+1}}$ *and* $\Delta_{[s_k,t+1]} > \sqrt{\frac{KC_0}{t-s_k+2}}$ **then**
    Let $e_k \leftarrow t, \mathcal{I}_k \leftarrow [s_k, e_k]$,
    $k \leftarrow k+1$, $s_k \leftarrow t+1$.
  end
  $t \leftarrow t+1$.
end
if $s_k \leq e$ **then**
  $e_k \leftarrow e, \mathcal{I}_k \leftarrow [s_k, e_k]$.
end

It is clear that this procedure ensures $\Delta_{\mathcal{I}_k} \leq \alpha_{\mathcal{I}_k}$ for all k. It remains to bound Γ. If $\Gamma > 1$, by the procedure and the decomposability of variation we have

$$\Delta_{[s,e]} \geq \Delta_{[s_1,e_1+1]} + \Delta_{[s_2,e_2+1]} + \ldots + \Delta_{[s_{\Gamma-1},e_{\Gamma-1}+1]} \geq \sum_{k=1}^{\Gamma-1}\sqrt{\frac{KC_0}{e_k - s_k + 2}} = \sum_{k=1}^{\Gamma-1}\sqrt{\frac{KC_0}{|\mathcal{I}_k|+1}}.$$

On the other hand Hölder's inequality implies

$$\left(\sum_{k=1}^{\Gamma-1}\sqrt{\frac{KC_0}{|\mathcal{I}_k|+1}}\right)^{\frac{2}{3}}\left(\sum_{k=1}^{\Gamma-1}(|\mathcal{I}_k|+1)\right)^{\frac{1}{3}}\geq(\Gamma-1)(KC_0)^{\frac{1}{3}}.$$

Combining the two inequalities above, we get

$$\Gamma-1\leq(KC_0)^{-\frac{1}{3}}\left(\sum_{k=1}^{\Gamma-1}(|\mathcal{I}_k|+1)\right)^{\frac{1}{3}}\Delta_{[s,e]}^{\frac{2}{3}}\leq\mathcal{O}\left((KC_0)^{-\frac{1}{3}}|\mathcal{J}|^{\frac{1}{3}}\Delta_{[s,e]}^{\frac{2}{3}}\right).$$

It is also clear from the condition $\Delta_{[s_k,t+1]}>\sqrt{\frac{KC_0}{t-s_k+2}}$ that the procedure creates one interval only when the distribution switches. Therefore, $\Gamma-1\leq S_{[s,e]}-1$, which completes the proof. ■

Appendix B. Concentration Results

This section is dedicated to all concentration results we need for our analysis. First we introduce some notations and technical lemmas.

Definition 10 *Define $U_t(\pi)$ as the conditional variance of the reward estimation for policy π at time t (given everything before time t), that is,*

$$U_t(\pi)=\mathbb{E}_t\left[(\widehat{r}_t(\pi(x_t))-\mathcal{R}_t(\pi))^2\right].$$

Also recall the variance notation defined in Section 3:

$$\widehat{V}_{\mathcal{I}}(Q,\nu,\pi)=\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\left[\frac{1}{Q^{\nu}(\pi(x_t)|x_t)}\right],\quad V_{\mathcal{I}}(Q,\nu,\pi)=\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\mathbb{E}_{x\sim\mathcal{D}_t^{\mathcal{X}}}\left[\frac{1}{Q^{\nu}(\pi(x)|x)}\right]$$

($\widehat{V}_t$ and V_t are shorthands for $\widehat{V}_{[t,t]}$ and $V_{[t,t]}$ respectively). The following lemma connects these notions of variance for our algorithm.

Lemma 11 *For any policy π and any time t in epoch i and block j, if $M_t\neq\emptyset$, then $U_t(\pi)\leq V_t(Q_{(i,m)},\nu_m,\pi)\log_2 T$ for any $m\in M_t$; if $M_t=\emptyset$, then $U_t(\pi)\leq V_t(Q_{(i,j)},\nu_j,\pi)$.*

Proof When $M_t\neq\emptyset$, the distribution over actions played by the algorithm is

$$p_t(\cdot)=\frac{1}{|M_t|}\sum_{m\in M_t}Q_{(i,m)}^{\nu_m}(\cdot|x_t).$$

Thus the variance is bounded as

$$\begin{aligned}U_t(\pi)&\leq\mathbb{E}\left[\widehat{r}_t(\pi(x_t))^2\right]=\mathbb{E}_{(x,r)\sim\mathcal{D}_t}\left[p_t(\pi(x))\cdot\frac{r_t(\pi(x))^2}{p_t(\pi(x))^2}\right]\\&\leq\mathbb{E}_{x\sim\mathcal{D}_t^{\mathcal{X}}}\left[\frac{|M_t|}{\sum_{m'\in M_t}Q_{(i,m')}^{\nu_{m'}}(\pi(x)|x)}\right]\leq\mathbb{E}_{x\sim\mathcal{D}_t^{\mathcal{X}}}\left[\frac{\log_2 T}{Q_{(i,m)}^{\nu_m}(\pi(x)|x)}\right]=V_t(Q_{(i,m)},\nu_m,\pi)\log_2 T,\end{aligned}$$

where we use the fact that $|M_t| \le j \le \log_2 \frac{T}{L} + 1 \le \log_2 T$. Similarly, when $M_t = \emptyset$, we have

$$U_t(\pi) \le \mathbb{E}_{x \sim \mathcal{D}_t^{\mathcal{X}}} \left[\frac{1}{Q_{(i,j)}^{\nu_j}(\pi(x)|x)} \right] = V_t(Q_{(i,j)}, \nu_j, \pi).$$

■

We repeatedly make use of the following standard concentration bounds for martingales (which is a version of the Freedman's inequality; see for example (Beygelzimer et al., 2011)).

Lemma 12 (Freedman's inequality) *Let $X_1, \ldots, X_n \in \mathbb{R}$ be a martingale difference sequence with respect to some filtration $\mathcal{F}_0, \mathcal{F}_1, \ldots$. Assume $X_i \le R$ a.s. for all i. Then for any $\delta \in (0,1)$ and $\lambda \in [0, 1/R]$, with probability at least $1-\delta$, we have*

$$\sum_{i=1}^{n} X_i \le \lambda V + \frac{\ln(1/\delta)}{\lambda}$$

where $V = \sum_{i=1}^{n} \mathbb{E}[X_i^2 | \mathcal{F}_{i-1}]$.

B.1. Concentration of reward estimator and variance

The following concentration results on the reward estimator and variance are based on applications of Lemma 12. Recall C_0 is defined in the algorithm.

Lemma 13 *With probability at least $1-\delta/4$, for all $Q \in \Delta^{\Pi}$, all $\nu \in \{\nu_0, \nu_1, \ldots, \nu_{j_{\max}}\}$, where $j_{\max} \triangleq \lceil \log_2 T \rceil$, all $\pi \in \Pi$, and all intervals $\mathcal{I}$, it holds that*

$$V_{\mathcal{I}}(Q,\nu,\pi) \le 6.4 \widehat{V}_{\mathcal{I}}(Q,\nu,\pi) + \frac{80 C_0}{\nu^2 |\mathcal{I}|}, \qquad \widehat{V}_{\mathcal{I}}(Q,\nu,\pi) \le 6.4 V_{\mathcal{I}}(Q,\nu,\pi) + \frac{80 C_0}{\nu^2 |\mathcal{I}|}. \tag{13}$$

Proof This is a consequence of the contexts being drawn independently, and is not related to the algorithm. Therefore, we can apply the same argument of (Dudík et al., 2011, Theorem 6), (Agarwal et al., 2014, Lemma 10), or (Luo et al., 2018, Lemma 15). For example, as shown by (Agarwal et al., 2014, Lemma 10), with probability $1 - \delta/(4T)$, for all Q, all π, and all $\mathcal{I}$, $V_{\mathcal{I}}(Q,\nu,\pi) - 6.4\widehat{V}_{\mathcal{I}}(Q,\nu,\pi)$ and $\widehat{V}_{\mathcal{I}}(Q,\nu,\pi) - 6.4 V_{\mathcal{I}}(Q,\nu,\pi)$ are both upper bounded by

$$\frac{75 \ln(|\Pi|)}{\nu^2 |\mathcal{I}|} + \frac{6.3 \ln(8T^3 |\Pi|^2/\delta)}{\nu |\mathcal{I}|} \le \frac{80 \ln(8T^3|\Pi|^2/\delta)}{\nu^2 |\mathcal{I}|}.$$

Another union bound over ν finishes the proof. ■

Lemma 14 *With probability at least $1-\delta/4$, for all policy $\pi \in \Pi$, we have for all interval $\mathcal{B}_j$ that corresponds to the first $j+1$ blocks of some epoch,*

$$\left| \widehat{\mathcal{R}}_{\mathcal{B}_j}(\pi) - \mathcal{R}_{\mathcal{B}_j}(\pi) \right| \le \frac{\nu_j}{|\mathcal{B}_j| \log_2 T} \sum_{t \in \mathcal{B}_j} U_t(\pi) + \frac{C_0 \log_2 T}{\nu_j |\mathcal{B}_j|},$$

and for all interval $\mathcal{A}$ that is covered by some replay phase of index m,

$$\left| \widehat{\mathcal{R}}_{\mathcal{A}}(\pi) - \mathcal{R}_{\mathcal{A}}(\pi) \right| \le \frac{\nu_m}{|\mathcal{A}| \log_2 T} \sum_{t \in \mathcal{A}} U_t(\pi) + \frac{C_0 \log_2 T}{\nu_m |\mathcal{A}|}.$$

Proof We simply apply Lemma 12 to the sequence $(\widehat{r}_t(\pi(x_t)) - \mathcal{R}_t(\pi))\mathbf{1}\{t \in \mathcal{B}_j\}$ for every interval and every π (with a union bound). Note that these random variables are bounded by $1/\nu_j$ so we can pick $\lambda = \nu_j/\log_2 T$. Similarly for the second statement we apply Lemma 12 to the sequence$(\widehat{r}_t(\pi(x_t)) - \mathcal{R}_t(\pi))\mathbf{1}\{m \in M_t\}$ with $\lambda = \nu_m/\log_2 T$. ■

Since most analysis conditions on these concentration results, we denote the event formally below, which clearly happens with probability at least $1 - \delta/2$.

Definition 15 (EVENT$_1$) *Define* **EVENT$_1$** *as the event that all bounds described in Lemma 14 and Lemma 13 hold.*

B.2. Concentration of regret

In this section we prove three main concentration results on regret, which play a crucial role later in our analysis. We focus on a specific epoch i and for simplicity use $\mathcal{B}_j$ and Q_j as shorthands for $\mathcal{B}_{(i,j)}$ and $Q_{(i,j)}$ respectively (we remind the reader that these notations are defined in the algorithm).

Lemma 16 *Assume* **EVENT$_1$** *holds, and assume that there is no restart triggered in $\mathcal{B}_j$, then the following hold for all $\pi \in \Pi$:*

$$\mathrm{Reg}_{\mathcal{B}_j}(\pi) \le 2\widehat{\mathrm{Reg}}_{\mathcal{B}_j}(\pi) + C_1\overline{K}\nu_j + C_2\Delta_{\mathcal{B}_j},$$
$$\widehat{\mathrm{Reg}}_{\mathcal{B}_j}(\pi) \le 2\mathrm{Reg}_{\mathcal{B}_j}(\pi) + C_1\overline{K}\nu_j + C_2\Delta_{\mathcal{B}_j},$$

where $C_1 = 2000, C_2 = 24$.

Lemma 17 *Assume* **EVENT$_1$** *holds. Let $\mathcal{A}$ be a complete replay phase of index m (that is, $|\mathcal{A}| = 2^m L$). If for all π, Eq. (4) and Eq. (5) in* ENDOFREPLAYTEST *do not hold, then the following hold for all π:*

$$\mathrm{Reg}_{\mathcal{A}}(\pi) \le 2\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + C_3\overline{K}\nu_m + C_4\Delta_{\mathcal{A}},$$
$$\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) \le 2\mathrm{Reg}_{\mathcal{A}}(\pi) + C_3\overline{K}\nu_m + C_4\Delta_{\mathcal{A}},$$

where $C_3 = 2 \times 10^6, C_4 = 24$.

Lemma 18 *Assume* **EVENT$_1$** *holds. Let $\mathcal{A} = [s, e]$ be a complete replay phase of index m (thus $|\mathcal{A}| = 2^m L$). Then the following hold for all π:*

$$\mathrm{Reg}_{\mathcal{A}}(\pi) \le 2\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + C_5\overline{K}\nu_m + C_6\Delta_{[\tau_i,e]},$$
$$\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) \le 2\mathrm{Reg}_{\mathcal{A}}(\pi) + C_5\overline{K}\nu_m + C_6\Delta_{[\tau_i,e]},$$

where $C_5 = 2000, C_6 = 24$.

To prove these results, we first prove the following auxiliary lemma. Basically it shows that in an interval $\mathcal{I}$, if we can bound the instant variance of a policy π by some quantity that is proportional to the regret of π, then $\mathrm{Reg}_{\mathcal{I}}(\pi)$ and $\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi)$ are close.

Lemma 19 *Assume* **EVENT**$_1$ *holds. Consider an interval $\mathcal{I}$ that is either $\mathcal{B}_j$ or $\mathcal{A}$ as defined in Lemma 14, and let μ be $\nu_j/\log_2 T$ if $\mathcal{I}$ is $\mathcal{B}_j$ and $\nu_m/\log_2 T$ if $\mathcal{I}$ is $\mathcal{A}$. If either of the following conditions holds:*

$$U_t(\pi) \le \frac{\mathrm{Reg}_{\mathcal{I}}(\pi)}{3\mu} + Z, \qquad \forall t \in \mathcal{I}, \forall \pi \in \Pi,$$
$$U_t(\pi) \le \frac{\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi)}{3\mu} + Z, \qquad \forall t \in \mathcal{I}, \forall \pi \in \Pi,$$

for some $Z > 0$, then we have

$$\mathrm{Reg}_{\mathcal{I}}(\pi) \le 2\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) + 8\mu Z + \frac{8C_0}{\mu|\mathcal{I}|}, \quad \widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) \le 2\mathrm{Reg}_{\mathcal{I}}(\pi) + 8\mu Z + \frac{8C_0}{\mu|\mathcal{I}|}.$$

Proof of Lemma 19. Suppose we have $U_t(\pi) \le \frac{\mathrm{Reg}_{\mathcal{I}}(\pi)}{3\mu} + Z$ for all $t \in \mathcal{I}$ and $\pi \in \Pi$, then

$$\begin{aligned}
&\mathrm{Reg}_{\mathcal{I}}(\pi) - \widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) \\
&= \mathcal{R}_{\mathcal{I}}(\pi^*_{\mathcal{I}}) - \mathcal{R}_{\mathcal{I}}(\pi) - \widehat{\mathcal{R}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) + \widehat{\mathcal{R}}_{\mathcal{I}}(\pi) \\
&\le \left(\mathcal{R}_{\mathcal{I}}(\pi^*_{\mathcal{I}}) - \widehat{\mathcal{R}}_{\mathcal{I}}(\pi^*_{\mathcal{I}})\right) + \left(\widehat{\mathcal{R}}_{\mathcal{I}}(\pi) - \mathcal{R}_{\mathcal{I}}(\pi)\right) && \text{(optimality of } \widehat{\pi}_{\mathcal{I}}\text{)} \\
&\le \frac{\mu}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} (U_t(\pi^*_{\mathcal{I}}) + U_t(\pi)) + \frac{2C_0}{\mu|\mathcal{I}|} && \textbf{(EVENT}_1\textbf{)} \\
&\le \frac{1}{3}\mathrm{Reg}_{\mathcal{I}}(\pi^*_{\mathcal{I}}) + \frac{1}{3}\mathrm{Reg}_{\mathcal{I}}(\pi) + 2\mu Z + \frac{2C_0}{\mu|\mathcal{I}|}, \\
&= \frac{1}{3}\mathrm{Reg}_{\mathcal{I}}(\pi) + 2\mu Z + \frac{2C_0}{\mu|\mathcal{I}|}, && (\mathrm{Reg}_{\mathcal{I}}(\pi^*_{\mathcal{I}}) = 0)
\end{aligned}$$

which gives $\mathrm{Reg}_{\mathcal{I}}(\pi) \le \frac{3}{2}\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) + 3\mu Z + \frac{3C_0}{\mu|\mathcal{I}|} \le 2\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) + 8\mu Z + \frac{8C_0}{\mu|\mathcal{I}|}$, proving the first part of the lemma. On the other hand,

$$\begin{aligned}
&\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) - \mathrm{Reg}_{\mathcal{I}}(\pi) \\
&= \widehat{\mathcal{R}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) - \widehat{\mathcal{R}}_{\mathcal{I}}(\pi) - \mathcal{R}_{\mathcal{I}}(\pi^*_{\mathcal{I}}) + \mathcal{R}_{\mathcal{I}}(\pi) \\
&\le \left(\widehat{\mathcal{R}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) - \mathcal{R}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}})\right) + \left(\mathcal{R}_{\mathcal{I}}(\pi) - \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)\right) && \text{(optimality of } \pi^*_{\mathcal{I}}\text{)} \\
&\le \frac{\mu}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} (U_t(\widehat{\pi}_{\mathcal{I}}) + U_t(\pi)) + \frac{2C_0}{\mu|\mathcal{I}|} && \textbf{(EVENT}_1\textbf{)} \\
&\le \frac{1}{3}\mathrm{Reg}_{\mathcal{I}}(\pi) + \frac{1}{3}\mathrm{Reg}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) + 2\mu Z + \frac{2C_0}{\mu|\mathcal{I}|} \\
&\le \frac{1}{2}\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) + \frac{1}{2}\widehat{\mathrm{Reg}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) + 4\mu Z + \frac{4C_0}{\mu|\mathcal{I}|} \\
&= \frac{1}{2}\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) + 4\mu Z + \frac{4C_0}{\mu|\mathcal{I}|}. && (\widehat{\mathrm{Reg}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) = 0)
\end{aligned}$$

where in the second to last inequality we use the fact $\mathrm{Reg}_{\mathcal{I}}(\pi) \le \frac{3}{2}\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) + 3\mu Z + \frac{3C_0}{\mu|\mathcal{I}|}$ for all π, which we just obtained previously. The last inequality gives $\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) \le 2\mathrm{Reg}_{\mathcal{I}}(\pi) + 8\mu Z + \frac{8C_0}{\mu|\mathcal{I}|}$, proving the second part.

The proof under the second condition proceeds in the exact same way. ■

Now we are ready to prove the three lemmas. We will frequently use the following facts:

$$\frac{C_0}{\nu_j^2|\mathcal{B}_j|} = \frac{C_0}{\nu_j^2 2^j L} = K, \quad \nu_0 = \sqrt{\frac{C_0}{K\lceil 4KC_0\rceil}} \in \left[\frac{1}{4K}, \frac{1}{2K}\right].$$

Proof of Lemma 16. Assume $\textbf{EVENT}_1$ holds. We prove by induction on j. When $j = 0$, we have $\mathrm{Reg}_{\mathcal{B}_0}(\pi) \le 1 \le 4\overline{K}\nu_0$, and

$$\begin{aligned}
\widehat{\mathrm{Reg}}_{\mathcal{B}_0}(\pi) - \mathrm{Reg}_{\mathcal{B}_0}(\pi) &= \widehat{\mathcal{R}}_{\mathcal{B}_0}(\widehat{\pi}_{\mathcal{B}_0}) - \widehat{\mathcal{R}}_{\mathcal{B}_0}(\pi) - \mathcal{R}_{\mathcal{B}_0}(\pi^*_{\mathcal{B}_0}) + \mathcal{R}_{\mathcal{B}_0}(\pi) \\
&\le \widehat{\mathcal{R}}_{\mathcal{B}_0}(\widehat{\pi}_{\mathcal{B}_0}) - \mathcal{R}_{\mathcal{B}_0}(\widehat{\pi}_{\mathcal{B}_0}) - \widehat{\mathcal{R}}_{\mathcal{B}_0}(\pi) + \mathcal{R}_{\mathcal{B}_0}(\pi) && \text{(by the optimality of } \pi^*_{\mathcal{B}_0}) \\
&\le 2\left(\frac{\nu_0}{|\mathcal{B}_0|}\sum_{t\in\mathcal{B}_0}\frac{1}{\nu_0} + \frac{C_0}{\nu_0 L}\right) \le 4, && (\textbf{EVENT}_1)
\end{aligned}$$

and thus $\widehat{\mathrm{Reg}}_{\mathcal{B}_0}(\pi) \le 5 \le 20\overline{K}\nu_0$. Below we prove the inequalities for a general j, assuming that they hold for $\{0, \dots, j-1\}$. For all $t \in \mathcal{B}_j$, and all $m \in [1, j]$,

$$\begin{aligned}
V_t(Q_m, \nu_m, \pi) &\le V_{\mathcal{B}_{m-1}}(Q_m, \nu_m, \pi) + \frac{\Delta_{\mathcal{B}_j}}{\nu_m} && \text{(Lemma 9)} \\
&\le 6.4\widehat{V}_{\mathcal{B}_{m-1}}(Q_m, \nu_m, \pi) + \frac{80C_0}{\nu_m^2|\mathcal{B}_{m-1}|} + \frac{\Delta_{\mathcal{B}_j}}{\nu_m} && (\textbf{EVENT}_1) \\
&\le 6.4\left(2K + \frac{\widehat{\mathrm{Reg}}_{\mathcal{B}_{m-1}}(\pi)}{C\nu_m}\right) + \frac{80C_0}{\nu_m^2 2^{m-1}L} + \frac{\Delta_{\mathcal{B}_j}}{\nu_m} && \text{(Eq. (2))} \\
&\le 6.4\left(2K + \frac{2\mathrm{Reg}_{\mathcal{B}_{m-1}}(\pi) + C_1\overline{K}\nu_{m-1} + C_2\Delta_{\mathcal{B}_j}}{C\nu_m}\right) + 160K + \frac{\Delta_{\mathcal{B}_j}}{\nu_m} \\
& && \text{(By induction hypothesis)} \\
&\le \frac{\mathrm{Reg}_{\mathcal{B}_{m-1}}(\pi)}{3\nu_m} + C_7\overline{K} + \frac{2\Delta_{\mathcal{B}_j}}{\nu_m} && (\text{let } C_7 = \tfrac{\sqrt{2}C_1}{C} + 12.8 + 160) \\
&\le \frac{\mathrm{Reg}_{\mathcal{B}_j}(\pi)}{3\nu_m} + C_7\overline{K} + \frac{3\Delta_{\mathcal{B}_j}}{\nu_m} && \text{(Lemma 8)} \\
&\le \frac{\mathrm{Reg}_{\mathcal{B}_j}(\pi)}{3\nu_j} + C_7\overline{K} + \frac{3\Delta_{\mathcal{B}_j}}{\nu_j}. && (\nu_m \ge \nu_j)
\end{aligned} \tag{14}$$

Besides, when $j = 0$, $V_t(Q_0, \nu_0, \pi) \le \frac{1}{\nu_0} \le 4\overline{K}$. By Lemma 11, we always have $U_t(\pi) \le V_t(Q_m, \nu_m, \pi)\log_2 T$ for some $m \in [0, j]$. Therefore, $U_t(\pi) \le \left(\frac{\mathrm{Reg}_{\mathcal{B}_j}(\pi)}{3\nu_j} + C_7\overline{K} + \frac{3\Delta_{\mathcal{B}_j}}{\nu_j}\right)\log_2 T$. Using Lemma 19 with $Z = \left(C_7\overline{K} + \frac{3\Delta_{B_j}}{\nu_j}\right)\log_2 T$, we get the two desired inequalities. ■

Proof of Lemma 17. For all $t \in \mathcal{A}$,

$$
\begin{aligned}
V_t(Q_m, \nu_m, \pi) &\le V_{\mathcal{A}}(Q_m, \nu_m, \pi) + \frac{\Delta_{\mathcal{A}}}{\nu_m} && \text{(Lemma 9)} \\
&\le 6.4\widehat{V}_{\mathcal{A}}(Q_m, \nu_m, \pi) + \frac{80C_0}{\nu_m^2|\mathcal{A}|} + \frac{\Delta_{\mathcal{A}}}{\nu_m} && (\textbf{EVENT}_1) \\
&\le 6.4\left(41\hat{V}_{\mathcal{B}_{j-1}}(Q_m, \nu_m, \pi) + D_2K\right) + 80K + \frac{\Delta_{\mathcal{A}}}{\nu_m} && \text{(Eq. (5) does not hold)} \\
&\le 263\widehat{V}_{\mathcal{B}_{j-1}}(Q_m, \nu_m, \pi) + (6.4D_2 + 80)K + \frac{\Delta_{\mathcal{A}}}{\nu_m} \\
&\le 263(41\widehat{V}_{\mathcal{B}_{m-1}}(Q_m, \nu_m, \pi) + D_5K) + (6.4D_2 + 80)K + \frac{\Delta_{\mathcal{A}}}{\nu_m} \\
& && \text{(Eq. (8) does not hold)} \\
&\le \frac{\widehat{\text{Reg}}_{\mathcal{B}_{m-1}}(\pi)}{1000\nu_m} + C_8K + \frac{\Delta_{\mathcal{A}}}{\nu_m} && \text{(by Eq. (2))} \\
& && \text{(let } C_8 = 263 \times 41 \times 2 + 263 \times D_5 + 6.4D_2 + 80\text{)} \\
&\le \frac{4\widehat{\text{Reg}}_{\mathcal{B}_{j-1}}(\pi) + D_4\overline{K}\nu_m}{1000\nu_m} + C_8K + \frac{\Delta_{\mathcal{A}}}{\nu_m} && \text{(Eq. (7))} \\
&= \frac{\widehat{\text{Reg}}_{\mathcal{B}_{j-1}}(\pi)}{250\nu_m} + (C_8 + 0.001D_4)\,\overline{K} + \frac{\Delta_{\mathcal{A}}}{\nu_m} \\
&\le \frac{4\widehat{\text{Reg}}_{\mathcal{A}}(\pi) + D_1\overline{K}\nu_m}{250\nu_m} + (C_8 + 0.001D_4)\overline{K} + \frac{\Delta_{\mathcal{A}}}{\nu_m} && \text{(Eq. (4) does not hold)} \\
&\le \frac{\widehat{\text{Reg}}_{\mathcal{A}}(\pi)}{3\nu_m} + C_9\overline{K} + \frac{\Delta_{\mathcal{A}}}{\nu_m}. && \text{(let } C_9 = \tfrac{D_1}{250} + C_8 + 0.001D_4\text{)}
\end{aligned}
$$

Using $U_t(\pi) \le V_t(Q_m, \nu_m, \pi)\log_2 T$ (Lemma 11) and invoking Lemma 19 with

$$Z = \left(C_9\overline{K} + \frac{\Delta_{\mathcal{A}}}{\nu_m}\right)\log_2 T$$

finish the proof. ■

Proof of Lemma 18. For all $t \in \mathcal{A}$,

$$
\begin{aligned}
V_t(Q_m, \nu_m, \pi) &\le V_{\mathcal{B}_{m-1}}(Q_m, \nu_m, \pi) + \frac{\Delta_{[\tau_i, e]}}{\nu_m} && \text{(Lemma 9)} \\
&\le 6.4\widehat{V}_{\mathcal{B}_{m-1}}(Q_m, \nu_m, \pi) + \frac{80C_0}{\nu_m^2|\mathcal{B}_{m-1}|} + \frac{\Delta_{[\tau_i, e]}}{\nu_m} && (\textbf{EVENT}_1) \\
&\le 6.4\left(2K + \frac{\widehat{\text{Reg}}_{\mathcal{B}_{m-1}}(\pi)}{C\nu_m}\right) + 160K + \frac{\Delta_{[\tau_i, e]}}{\nu_m} && \text{(Eq. (2))} \\
&\le \frac{6.4\left(2\text{Reg}_{\mathcal{B}_{m-1}}(\pi) + C_1\overline{K}\nu_{m-1} + C_2\Delta_{\mathcal{B}_{m-1}}\right)}{C\nu_m} + (12.8K + 160K) + \frac{\Delta_{[\tau_i, e]}}{\nu_m} \\
& && \text{(Lemma 16)}
\end{aligned}
$$

$$\leq \frac{\mathrm{Reg}_{\mathcal{A}}(\pi)}{3\nu_m} + C_{10}\overline{K} + \frac{2\Delta_{[\tau_i,e]}}{\nu_m}. \qquad \text{(Lemma 8)}$$

$$\text{(let } C_{10} = \tfrac{6.4\sqrt{2}C_1}{C} + 12.8 + 160\text{)}$$

Using $U_t(\pi) \leq V_t(Q_m, \nu_m, \pi)\log_2 T$ by Lemma 11 and invoking Lemma 19 with

$$Z = \left(C_{10}\overline{K} + \frac{2\Delta_{[\tau_i,e]}}{\nu_m}\right)\log_2 T$$

finish the proof. ■

Appendix C. Omitted Details in Section 5.2 – Bounding Individual Regret Terms

In Section 5.2, we have partitioned a block $\mathcal{J}$ into $\Gamma = \mathcal{O}\left(\min\{S_{\mathcal{J}}, 1 + (KC_0)^{-1/3}\Delta_{\mathcal{J}}^{2/3}|\mathcal{J}|^{1/3}\}\right)$ intervals $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_\Gamma$, such that each one has $\Delta_{\mathcal{I}_k} \leq \alpha_{\mathcal{I}_k}$. In particular, we use the procedure described in Algorithm 2, which only happens in the analysis, to do the partition. Then we obtain a regret bound of block $\mathcal{J}$ up to the first $\Gamma - 1$ intervals as in Eq. (12).

For the remaining interval $\mathcal{I}_\Gamma$, because it might be interupted by restart, the terms $\alpha_{\mathcal{I}_\Gamma}$ and $\varepsilon_{\mathcal{I}_\Gamma}$ produced by Lemma 4 would depend on when we end the block, which is random and makes the analysis difficult. We resolve this issue by introducing the following *fictitious block* and a new partition over it.

Definition 20 (fictitious block) *Define*

$$\mathcal{J}' \triangleq [\tau_i + 2^{j-1}L, \tau_i + 2^j L - 1], \tag{15}$$

and let $\mathcal{I}'_1 \cup \mathcal{I}'_2, \ldots \cup \mathcal{I}'_\Upsilon$ be a partition of $\mathcal{J}'$ using the procedure in Algorithm 2.

Comparing the definition of $\mathcal{J}'$ to that of $\mathcal{J}$ in Eq. (11), one can see that $\mathcal{J}$ and $\mathcal{J}'$ only differ when there is a restart triggered in block j. Put differently, $\mathcal{J}'$ is the *planned* block j while $\mathcal{J}$ is the realized block j. Conditioned on all history before block j, $\mathcal{J}'$, as well as the intervals $\mathcal{I}'_1, \ldots, \mathcal{I}'_\Upsilon$ and the excess regret and excess regret thresholds defined on them, are determined, while $\mathcal{J}, \mathcal{I}_1, \ldots, \mathcal{I}_\Gamma$ and similar quantities on them are random. The following facts are clear by the procedure in Algorithm 2:

Fact 21 *Let $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_\Gamma\}$ be the partition of $\mathcal{J}$ (defined in (11)) using Algorithm 2, and also $\{\mathcal{I}'_1, \mathcal{I}'_2, \ldots, \mathcal{I}'_\Upsilon\}$ be the partition of $\mathcal{J}'$ (defined in (15)) using the same algorithm. Then (a) $\Gamma \leq \Upsilon$, (b) $\mathcal{I}_k = \mathcal{I}'_k, \forall k \in [\Gamma - 1]$, and (c) $s_{\mathcal{I}_\Gamma} = s_{\mathcal{I}'_\Gamma}, e_{\mathcal{I}_\Gamma} \leq e_{\mathcal{I}'_\Gamma}$, where $\mathcal{I}_\Gamma \triangleq [s_{\mathcal{I}_\Gamma}, e_{\mathcal{I}_\Gamma}], \mathcal{I}'_\Gamma \triangleq [s_{\mathcal{I}'_\Gamma}, e_{\mathcal{I}'_\Gamma}]$.*

With the above new definitions, we have the following specialized lemma for the regret in $\mathcal{I}_\Gamma$, which is an analogue of Lemma 4:

Lemma 22 *With probability $1 - \delta$, ADA-ILTCB$^+$ guarantees the following for $\mathcal{I} = \mathcal{I}_\Gamma$ and $\mathcal{I}' = \mathcal{I}'_\Gamma$ (recall j is the index of the block that contains $\mathcal{I}_\Gamma$):*

$$\sum_{t\in\mathcal{I}} (r_t(\pi_t^*(x_t)) - r_t(a_t)) \leq \mathcal{O}\left(\left(\sum_{t\in\mathcal{I}}\sum_{m\in M_t\cup\{j\}} \overline{K}\nu_m\right) + |\mathcal{I}|\alpha_{\mathcal{I}'} + |\mathcal{I}|\Delta_{\mathcal{I}'} + |\mathcal{I}|\varepsilon_{\mathcal{I}'}\mathbf{1}\{\varepsilon_{\mathcal{I}'} > D_3\alpha_{\mathcal{I}'}\}\right). \tag{16}$$

Proof The proof is the same as Lemma 4, except that we bound $\sum_{\pi\in\Pi} Q_m(\pi)\mathrm{Reg}_t(\pi)$ slightly differently:

$$\begin{aligned}
\sum_{\pi\in\Pi} Q_m(\pi)\mathrm{Reg}_t(\pi) &\le \sum_{\pi\in\Pi} Q_m(\pi)\mathrm{Reg}_{\mathcal{I}'}(\pi) + \mathcal{O}(\Delta_{\mathcal{I}'}) && \text{(Lemma 8)}\\
&= \sum_{\pi\in\Pi} 8Q_m(\pi)\widehat{\mathrm{Reg}}_{\mathcal{B}_{j-1}}(\pi) + \mathcal{O}(\Delta_{\mathcal{I}'}) + \varepsilon_{\mathcal{I}'} && \text{(definition of } \varepsilon_{\mathcal{I}'})\\
&\le \sum_{\pi\in\Pi} 8Q_m(\pi)\left(4\widehat{\mathrm{Reg}}_{\mathcal{B}_{m-1}}(\pi) + D_4\overline{K}\nu_m\right) + \mathcal{O}(\Delta_{\mathcal{I}'}) + \varepsilon_{\mathcal{I}'}\\
& && \text{(Eq. (6) does not hold for block } j-1)\\
&\le \mathcal{O}(\overline{K}\nu_m + \Delta_{\mathcal{I}'}) + \varepsilon_{\mathcal{I}'} && \text{(Eq. (1))}\\
&\le \mathcal{O}(\overline{K}\nu_m + \alpha_{\mathcal{I}'} + \Delta_{\mathcal{I}'}) + \varepsilon_{\mathcal{I}'}\mathbf{1}\left\{\varepsilon_{\mathcal{I}'} > D_3\alpha_{\mathcal{I}'}\right\}.
\end{aligned}$$

Combining this with the rest of the proof of Lemma 4 finishes the proof. ■

Using the this lemma, we write the regret in $\mathcal{I}'_\Gamma$ as three terms similar to those in Eq. (12):

$$\underbrace{\sum_{t\in\mathcal{I}_\Gamma}\sum_{m\in M_t\cup\{j\}}\mathcal{O}(\overline{K}\nu_m)}_{\overline{\text{TERM}}_1} + \underbrace{\mathcal{O}(|\mathcal{I}_\Gamma|\alpha_{\mathcal{I}'_\Gamma})}_{\overline{\text{TERM}}_2} + \underbrace{\mathcal{O}(|\mathcal{I}_\Gamma|\varepsilon_{\mathcal{I}'_\Gamma}\mathbf{1}\{\varepsilon_{\mathcal{I}'_\Gamma} > D_3\alpha_{\mathcal{I}'_\Gamma}\})}_{\overline{\text{TERM}}_3}. \tag{17}$$

The rest of this section bounds the three terms $\text{TERM}_1+\overline{\text{TERM}}_1$, $\text{TERM}_2+\overline{\text{TERM}}_2$, and $\text{TERM}_3+\overline{\text{TERM}}_3$ separately. Combining these bounds proves Lemma 6.

Lemma 23 (Bounding $\text{TERM}_1 + \overline{\text{TERM}}_1$) *With probability at least $1-\delta/4$, for all block $\mathcal{J}$ with index j,*

$$\text{TERM}_1 + \overline{\text{TERM}}_1 = \sum_{k=1}^{\Gamma}\sum_{t\in\mathcal{I}_k}\sum_{m\in M_t\cup\{j\}}\overline{K}\nu_m \le \widetilde{\mathcal{O}}\left(\log(1/\delta)\sqrt{KC_0 2^j L}\right).$$

Proof Recall that $\mathcal{J} = \mathcal{I}_1\cup\cdots\cup\mathcal{I}_\Gamma$, and that there is no more than $2^j L$ steps in block $\mathcal{J}$. At every step with probability at most $\frac{1}{L}2^{-j/2}2^{-m/2}$ we start a replay interval with length $2^m L$. Therefore, with $\mathbb{I}_{t,m} \triangleq \mathbf{1}\{\text{the algorithm starts a replay phase of index } m \text{ at time } t\}$ we have

$$\begin{aligned}
\sum_{t\in\mathcal{J}}\sum_{m\in M_t\cup\{j\}}\overline{K}\nu_m &\le 2^j L\times\overline{K}\nu_j + \sum_{m=0}^{j-1}\sum_{t\in\mathcal{J}}\mathbb{I}_{t,m}\times 2^m L\times\overline{K}\nu_m\\
&= \log_2 T\left(\sqrt{KC_0 2^j L} + \sum_{m=0}^{j-1}\sum_{t\in\mathcal{J}}\mathbb{I}_{t,m}\sqrt{KC_0 2^m L}\right). \qquad (18)
\end{aligned}$$

Note that $\sum_{t\in\mathcal{J}}\mathbb{E}_t[\mathbb{I}_{t,m}] \le 2^j L\times\frac{1}{L}2^{-j/2}2^{-m/2} \le 2^{\frac{j-m}{2}}$. By Freedman's inequality (Lemma 12 with $\lambda=1$) and a union bound over all possible $\mathcal{J}$, all j's and all m's, we have that with probability at least $1-\delta/4$, for all possible $\mathcal{J}$ and j and m,

$$\sum_{t\in\mathcal{J}}\mathbb{I}_{t,m} \le \sum_{t\in\mathcal{J}}\mathbb{E}_t[\mathbb{I}_{t,m}] + \sum_{t\in\mathcal{J}}\mathbb{E}_t[\mathbb{I}^2_{t,m}] + \log\frac{4T^3}{\delta}$$

$$\leq 2^{\frac{j-m}{2}+1} + \log \frac{4T^3}{\delta}$$

Combining this with Eq. (18) and noting $j \leq \log_2 T$, we get

$$\sum_{t \in \mathcal{J}} \sum_{m \in M_t \cup \{j\}} \overline{K} \nu_m \leq \log_2 T \left(\sqrt{KC_0 2^j L} + (\log_2 T) \mathcal{O}\left(\sqrt{KC_0 2^j L} \log(1/\delta) \right) \right)$$
$$= \widetilde{\mathcal{O}}\left(\log(1/\delta) \sqrt{KC_0 2^j L} \right).$$

■

Lemma 24 (Bounding $\textsc{Term}_2 + \overline{\textsc{Term}}_2$) *For all block $\mathcal{J}$,*

$$\textsc{Term}_2 + \overline{\textsc{Term}}_2 = \left(\sum_{k=1}^{\Gamma-1} |\mathcal{I}_k| \alpha_{\mathcal{I}_k} \right) + |\mathcal{I}_\Gamma| \alpha_{\mathcal{I}'_\Gamma}$$
$$\leq \mathcal{O}\left(\log_2 T \times \min\left\{ \sqrt{KC_0 S_{\mathcal{J}} |\mathcal{J}|}, \sqrt{KC_0|\mathcal{J}|} + (KC_0)^{\frac{1}{3}} (\Delta_{\mathcal{J}})^{\frac{1}{3}} |\mathcal{J}|^{\frac{2}{3}} \right\} \right).$$

Proof Note $\alpha_{\mathcal{I}'_\Gamma} \leq \alpha_{\mathcal{I}_\Gamma}$ (because $\mathcal{I}_\Gamma \subseteq \mathcal{I}'_\Gamma$). Therefore the left hand side is upper bounded by $\sum_{k=1}^{\Gamma} |\mathcal{I}_k| \alpha_{\mathcal{I}_k}$. We simply plug in the definition of $\alpha_{\mathcal{I}_k}$, apply Cauchy-Schwarz inequality, and use the bound on Γ from Lemma 5 to get:

$$\sum_{k=1}^{\Gamma} |\mathcal{I}_k| \alpha_{\mathcal{I}_k} = \log_2 T \times \sum_{k=1}^{\Gamma} \sqrt{2KC_0 |\mathcal{I}_k|} \leq \log_2 T \times \sqrt{2KC_0 \Gamma |\mathcal{J}|}$$
$$\leq \mathcal{O}\left(\log_2 T \times \min\left\{ \sqrt{KC_0 S_{\mathcal{J}} |\mathcal{J}|}, \sqrt{KC_0|\mathcal{J}|} + (KC_0)^{\frac{1}{3}} (\Delta_{\mathcal{J}})^{\frac{1}{3}} |\mathcal{J}|^{\frac{2}{3}} \right\} \right).$$

■

C.1. Bounding $\textsc{Term}_3 + \overline{\textsc{Term}}_3$

The analysis of $\textsc{Term}_3 + \overline{\textsc{Term}}_3$ heavily relies on the definition of the fictitious block $\mathcal{J}'$ and the partition $\mathcal{I}'_1 \cup \cdots \cup \mathcal{I}'_\Upsilon$ on it, which are defined at the beginning of Appendix C.

For an interval $\mathcal{I} \subseteq \mathcal{J}'$, $\textsc{Term}_3$ only contributes to regret when the interval satisfies $\varepsilon_{\mathcal{I}} > D_3 \alpha_{\mathcal{I}}$. These intervals have large *excess regret* that causes extra regret. However, we will argue that the larger the excess regret, the sooner the algorithm can detect the non-stationarity and restart the algorithm. To prove this, we make use of the following lemma.

Lemma 25 *Assume $\textsc{Event}_1$ holds. Let $\mathcal{I} = [s, e]$ be an interval in the fictitious block $\mathcal{J}'$ with index j, and such that $\Delta_{\mathcal{I}} \leq \alpha_{\mathcal{I}}$ and $\varepsilon_{\mathcal{I}} > D_3 \alpha_{\mathcal{I}}$. Then*

(a) there exists an index $m_{\mathcal{I}} \in \{0, 1, \ldots, j-1\}$ such that $D_3 \overline{K} \nu_{m+1} < \varepsilon_{\mathcal{I}} \leq D_3 \overline{K} \nu_m$;

(b) $|\mathcal{I}| > 2^{m_{\mathcal{I}}} L$;

(c) if the algorithm starts a replay phase $\mathcal{A}$ with index $m_{\mathcal{I}}$ within the range of $[s, e - 2^{m_{\mathcal{I}}} L]$, then the algorithm restarts when the replay phase finishes.

Proof For notation simplicity, we use m as shorthand for $m_{\mathcal{I}}$. For (a), simply note that on one hand $\varepsilon_{\mathcal{I}} \le \max_\pi \mathrm{Reg}_{\mathcal{I}}(\pi) \le 1 \le D_3\overline{K}\nu_0$; and on the other hand, $\varepsilon_{\mathcal{I}} > D_3\alpha_{\mathcal{I}} = D_3\sqrt{\frac{2KC_0}{|\mathcal{I}|}}\log_2 T \ge D_3\sqrt{\frac{KC_0}{2^j L}}\log_2 T = D_3\overline{K}\nu_j$, where the second inequality is because $|\mathcal{I}| \le |\mathcal{J}'| \le 2^{j-1}L$.

For (b), note that $D_3\sqrt{\frac{2KC_0}{|\mathcal{I}|}}\log_2 T = D_3\alpha_{\mathcal{I}} < \varepsilon_{\mathcal{I}} \le D_3\overline{K}\nu_m = D_3\sqrt{\frac{KC_0}{2^m L}}\log_2 T$, which implies $|\mathcal{I}| > 2\times 2^m L$.

For (c), we show that the ENDOFREPLAYTEST fails when the replay phase finishes. That is, one of Eq. (3)-Eq. (5) will hold for some π. Suppose for all $\pi\in\Pi$, Eq. (4) and Eq. (5) do not hold, then by Lemma 17 we have for all π,

$$\begin{aligned}
\mathrm{Reg}_{\mathcal{A}}(\pi) &\le 2\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + C_3\overline{K}\nu_m + C_4\Delta_{\mathcal{A}} \\
&\le 2\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + C_3\overline{K}\nu_m + C_4\Delta_{\mathcal{I}} && (\mathcal{A}\text{ lies in }\mathcal{I}) \\
&\le 2\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + C_3\overline{K}\nu_m + C_4\alpha_{\mathcal{I}} && (\text{by the condition}) \\
&\le 2\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + C_3\overline{K}\nu_m + C_4\overline{K}\nu_m && (D_3\alpha_{\mathcal{I}} < \varepsilon_{\mathcal{I}} \le D_3\overline{K}\nu_m) \\
&= 2\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + (C_3+C_4)\overline{K}\nu_m.
\end{aligned}$$

Also by the definition of excess regret, there exists π' such that

$$\begin{aligned}
\mathrm{Reg}_{\mathcal{A}}(\pi') &\ge \mathrm{Reg}_{\mathcal{I}}(\pi') - 2\Delta_{\mathcal{I}} && (\text{Lemma 8}) \\
&\ge 8\widehat{\mathrm{Reg}}_{\mathcal{B}_{j-1}}(\pi') + \varepsilon_{\mathcal{I}} - 2\alpha_{\mathcal{I}} \\
&\ge 8\widehat{\mathrm{Reg}}_{\mathcal{B}_{j-1}}(\pi') + D_3\overline{K}\nu_{m+1} - 2\overline{K}\nu_m \\
&\ge 8\widehat{\mathrm{Reg}}_{\mathcal{B}_{j-1}}(\pi') + (0.5D_3 - 2)\overline{K}\nu_m.
\end{aligned}$$

Combining the two inequalities we get

$$\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi') > 4\widehat{\mathrm{Reg}}_{B_{j-1}}(\pi') + \frac{0.5D_3 - 2 - C_3 - C_4}{2}\overline{K}\nu_m > 4\widehat{\mathrm{Reg}}_{B_{j-1}}(\pi') + D_1\overline{K}\nu_m,$$

which makes Eq. (3) hold. ■

Now we are ready to bound $\mathrm{TERM}_3 + \overline{\mathrm{TERM}_3}$:

Lemma 26 (Bounding $\mathrm{TERM}_3 + \overline{\mathrm{TERM}_3}$) *With probability at least $1-\delta$,*

$$\begin{aligned}
\mathrm{TERM}_3 + \overline{\mathrm{TERM}_3} &= \sum_{k=1}^{\Gamma-1} |\mathcal{I}_k|\varepsilon_{\mathcal{I}_k}\mathbf{1}\{\varepsilon_{\mathcal{I}_k} > D_3\alpha_{\mathcal{I}_k}\} + |\mathcal{I}_\Gamma|\varepsilon_{\mathcal{I}'_\Gamma}\mathbf{1}\{\varepsilon_{\mathcal{I}'_\Gamma} > D_3\alpha_{\mathcal{I}'_\Gamma}\} \\
&\le \mathcal{O}\left(\log(1/\delta)\log(T)\sqrt{KC_0\Gamma 2^j L}\right) \\
&\le \mathcal{O}\left(\log(1/\delta)\log(T)\min\left\{\sqrt{KC_0 S_{\mathcal{J}} 2^j L}, \sqrt{KC_0 2^j L} + (KC_0)^{\frac{1}{3}}\Delta_{\mathcal{J}}^{\frac{1}{3}}(2^j L)^{\frac{2}{3}}\right\}\right).
\end{aligned}$$

Proof We condition on the history before block j. As discussed at the beginning of Appendix C, under this condition, the partition $\mathcal{I}'_1,\dots,\mathcal{I}'_\Upsilon$, as well as excess regret and excess regret thresholds

defined on them are all fixed. Define $\mathcal{K} = \{k \in [\Upsilon] \mid \varepsilon_{\mathcal{I}'_k} > D_3\alpha_{\mathcal{I}'_k}\}$. Then by Fact 21, $\textsc{Term}_3 + \overline{\textsc{Term}_3}$ can be written as

$$\sum_{k\in[\Gamma]} |\mathcal{I}_k|\varepsilon_{\mathcal{I}'_k}\mathbf{1}\{\varepsilon_{\mathcal{I}'_k} > D_3\alpha_{\mathcal{I}'_k}\} = \sum_{k\in[\Gamma]\cap\mathcal{K}} |\mathcal{I}_k|\varepsilon_{\mathcal{I}'_k}.$$

For $k \in \mathcal{K}$, denote by m_k the index $m_{\mathcal{I}'_k}$ defined in Lemma 25. Then

$$\begin{aligned}
\sum_{k\in[\Gamma]\cap\mathcal{K}} |\mathcal{I}_k|\varepsilon_{\mathcal{I}'_k} &\le \sum_{k\in[\Gamma]\cap\mathcal{K}} |\mathcal{I}_k| D_3 K\nu_{m_k} \\
&= \sum_{k\in[\Gamma]\cap\mathcal{K}} \left(2^{m_k}L \times D_3\overline{K}\nu_{m_k} + (|\mathcal{I}_k| - 2^{m_k}L) \times D_3\overline{K}\nu_{m_k}\right) \\
&\le \underbrace{\sum_{k\in[\Gamma]\cap\mathcal{K}} (\log_2 T) D_3\sqrt{KC_0 2^{m_k}L}}_{\textsc{Term}_4} + \underbrace{\sum_{k\in[\Gamma]\cap\mathcal{K}} (|\mathcal{I}_k| - 2^{m_k}L) D_3\overline{K}\nu_{m_k}}_{\textsc{Term}_5}.
\end{aligned}$$

$\textsc{Term}_4$ can be bounded as

$$\begin{aligned}
\sum_{k\in[\Gamma]\cap\mathcal{K}} (\log_2 T) D_3\sqrt{KC_0 2^{m_k}L} &\le \sum_{k\in[\Gamma]\cap\mathcal{K}} (\log_2 T) D_3\sqrt{KC_0|\mathcal{I}'_k|} && \text{(Lemma 25, (b))} \\
&\le (\log_2 T) D_3\sqrt{KC_0\Gamma \sum_{k\in[\Gamma]\cap\mathcal{K}} |\mathcal{I}'_k|} && \text{(Cauchy-Schwarz)} \\
&= \mathcal{O}\left((\log T)\sqrt{KC_0\Gamma 2^j L}\right).
\end{aligned}$$

Next we want to bound $\textsc{Term}_5$. Rewrite it as the following (denote $\mathcal{I}_k = [s_k, e_k], \mathcal{I}'_k = [s'_k, e'_k]$):

$$\begin{aligned}
\textsc{Term}_5 &\le \sum_{k\in[\Gamma]\cap\mathcal{K}} \sum_{t\in[s_k+2^{m_k}L, e_k]} D_3\overline{K}\nu_{m_k} = \sum_{k\in\mathcal{K}} \sum_{t\in[s_k+2^{m_k}L, e_k]} D_3\overline{K}\nu_{m_k}\mathbf{1}\{t \le e_\Gamma\} \\
&= \sum_{k\in\mathcal{K}} \sum_{t\in[s'_k+2^{m_k}L, e'_k]} D_3\overline{K}\nu_{m_k}\mathbf{1}\{t \le e_\Gamma\}
\end{aligned}$$

Below we show how to upper bound for a number z the probability $\Pr\{\textsc{Term}_5 > z\}$. Define the following function:

$$f(\tau) \triangleq \sum_{k\in\mathcal{K}} \sum_{t\in[s'_k+2^{m_k}L, e'_k]} D_3\overline{K}\nu_{m_k}\mathbf{1}\{t \le \tau\},$$

which is again not random conditioning on the history before block j. Our strategy is to first bound $\Pr\{\textsc{Term}_5 > f(\tau)\}$ by some function of $f(\tau)$. First by comparing definitions it is clear that $\Pr\{\textsc{Term}_5 > f(\tau)\} \le \Pr\{e_\Gamma > \tau\}$. Note that $e_\Gamma > \tau$ is the event that the block ends later than τ. From Lemma 25, we know that in interval $\mathcal{I}'_k$ with $k \in \mathcal{K}$, if the algorithm starts an replay phase at some $t^* \in [s_k, e_k - 2^{m_k}L]$, then conditioned on $\textbf{EVENT}_1$, the algorithm will restart at (or before) $t^* + 2^{m_k}L$. That is, for an intervals $k \in \mathcal{K}$, the algorithm always has $|\mathcal{I}'_k| - 2^{m_k}L$ opportunities to start a replay phase with index m_k which triggers restart eventually. Thus, if the algorithm proceeds

to time τ, and has not restarted yet, it has missed all such opportunities before τ. More precisely, for all $k \in \mathcal{K}$ with $e'_k < \tau$ (i.e., the intervals that are before τ), the algorithms misses all opportunities in

$$[s'_k, e'_k - 2^{m_k} L]$$

to start a replay phase with index m_k; for k such that $\tau \in [s'_k, e'_k]$ (i.e., the interval where τ lies in), the algorithm misses all opportunities in

$$[s'_k, \tau - 2^{m_k} L]$$

to start a replay phase with index m_k (define this set to be empty if $\tau - 2^{m_k} L < s'_k$). Combining these cases, we see that the probability that the block ends later than τ (i.e., $e_\Gamma > \tau$) is smaller than

$$\begin{aligned}
&\prod_{k \in \mathcal{K}} \prod_{t \in [s'_k, e'_k - 2^{m_k} L]} (1 - q_{m_k} \mathbf{1}\{t \le \tau - 2^{m_k} L\}), \\
&= \prod_{k \in \mathcal{K}} \prod_{t \in [s'_k + 2^{m_k} L, e'_k]} (1 - q_{m_k} \mathbf{1}\{t \le \tau\}).
\end{aligned}$$

where $q_m = \frac{1}{L}\sqrt{\frac{1}{2^j 2^m}} = \sqrt{\frac{K}{C_0 2^j L}}\nu_m$ is the probability to start a replay phase with index m at any time. Using the inequality $1 - x \le e^{-x}$, the above probability can further be upper bounded by

$$\begin{aligned}
&\prod_{k \in \mathcal{K}} \prod_{t \in [s'_k + 2^{m_k} L, e'_k]} \exp\left(-q_{m_k} \mathbf{1}\{t \le \tau\}\right) \\
&= \exp\left(-\sum_{k \in \mathcal{K}} \sum_{t \in [s'_k + 2^{m_k} L, e'_k]} q_{m_k} \mathbf{1}\{t \le \tau\}\right) \\
&= \exp\left(-\sum_{k \in \mathcal{K}} \sum_{t \in [s'_k + 2^{m_k} L, e'_k]} \sqrt{\frac{K}{C_0 2^j L}} \nu_{m_k} \mathbf{1}\{t \le \tau\}\right) \\
&= \exp\left(-\sum_{k \in \mathcal{K}} \sum_{t \in [s'_k + 2^{m_k} L, e'_k]} D_3 \overline{K} \nu_{m_k} \mathbf{1}\{t \le \tau\} \cdot \frac{1}{C^*}\right) \\
&= \exp\left(-\frac{f(\tau)}{C^*}\right),
\end{aligned}$$

where $C^* \triangleq (\log_2 T) D_3 \sqrt{K C_0 2^j L}$. Combining all arguments above, we have

$$\Pr\{\textsc{Term}_5 > f(\tau)\} \le \Pr\{e_\Gamma > \tau\} \le \exp\left(-\frac{f(\tau)}{C^*}\right).$$

Picking $z = \log(e/\delta) C^*$, we then let τ be such that $f(\tau) \le z < f(\tau + 1)$. If no such τ exists, $\Pr[\textsc{Term}_5 > z] = 0$; otherwise, since $f(\tau) \ge f(\tau + 1) - D_3(\log_2 T) > z - D_3(\log_2 T)$, we have

$$\Pr[\textsc{Term}_5 > z] \le \Pr[\textsc{Term}_5 > f(\tau)] \le \exp\left(-\frac{f(\tau)}{C^*}\right) \le \exp\left(\frac{D_3 \log_2 T}{C^*} - \frac{z}{C^*}\right)$$

$$\leq e \times \frac{\delta}{e} = \delta.$$

Therefore, we conclude that $\textsc{Term}_5 \geq \log(e/\delta)C^* = \log(e/\delta)(\log_2 T)D_3\sqrt{KC_0 2^j L}$ with probability at most δ. Combining $\textsc{Term}_4$ and $\textsc{Term}_5$, we get the first bound. Further using Lemma 5 gives the second bound. ■

Appendix D. Bounding the Number of Restarts

Lemma 27 *Assume* $\textbf{EVENT}_1$ *holds. Then for all* t *in epoch* i *with* $\Delta_{[\tau_i,t]} \leq \sqrt{\frac{KC_0}{t-\tau_i+1}}$, *restart will not be triggered at time* t.

Proof We will show that both tests pass and thus the algorithm does not restart.

No restarts by ENDOFBLOCKTEST. Let $t = \tau_i + 2^j L - 1$ for some j. Then $\Delta_{[\tau_i,t]} \leq \sqrt{\frac{KC_0}{t-\tau_i+1}} = K\nu_j$. For any $\pi \in \Pi, k \in [0, j-1]$,

$$\begin{aligned}
\widehat{\text{Reg}}_{\mathcal{B}_j} &\leq 2\text{Reg}_{\mathcal{B}_j} + C_1\overline{K}\nu_j + C_2\Delta_{[\tau_i,t]} && \text{(Lemma 16)}\\
&\leq 2\text{Reg}_{\mathcal{B}_k} + C_1\overline{K}\nu_j + (C_2+4)\Delta_{[\tau_i,t]} && \text{(Lemma 8)}\\
&\leq 2\left(2\widehat{\text{Reg}}_{\mathcal{B}_k} + C_1\overline{K}\nu_k + C_2\Delta_{[\tau_i,t]}\right) + C_1\overline{K}\nu_j + (C_2+4)\Delta_{[\tau_i,t]} && \text{(Lemma 16)}\\
&\leq 4\widehat{\text{Reg}}_{\mathcal{B}_k} + 3C_1\overline{K}\nu_k + (3C_2+4)\Delta_{[\tau_i,t]}\\
&\leq 4\widehat{\text{Reg}}_{\mathcal{B}_k} + D_4\overline{K}\nu_k. && (3C_1+3C_2+4 \leq D_4)
\end{aligned}$$

Similarly, $\widehat{\text{Reg}}_{\mathcal{B}_k} \leq 4\widehat{\text{Reg}}_{\mathcal{B}_j} + D_4\overline{K}\nu_k$. On the other hand,

$$\begin{aligned}
\widehat{V}_{\mathcal{B}_k}(Q_{k+1},\nu_{k+1},\pi) &\leq 6.4V_{\mathcal{B}_k}(Q_{k+1},\nu_{k+1},\pi) + \frac{80C_0}{\nu_{k+1}^2|\mathcal{B}_k|} && (\textbf{EVENT}_1)\\
&\leq 6.4V_{\mathcal{B}_j}(Q_{k+1},\nu_{k+1},\pi) + \frac{80C_0}{\nu_{k+1}^2|\mathcal{B}_k|} + \frac{6.4\Delta_{[\tau_i,t]}}{\nu_{k+1}} && \text{(Lemma 9)}\\
&\leq 6.4\left(6.4\widehat{V}_{\mathcal{B}_j}(Q_{k+1},\nu_{k+1},\pi) + \frac{80C_0}{\nu_{k+1}^2|\mathcal{B}_j|}\right) + \frac{80C_0}{\nu_{k+1}^2|\mathcal{B}_k|} + \frac{6.4\Delta_{[\tau_i,t]}}{\nu_{k+1}} && (\textbf{EVENT}_1)\\
&\leq 41\widehat{V}_{\mathcal{B}_j}(Q_{k+1},\nu_{k+1},\pi) + D_5K. && (6.4\times 80 + 160 + 6.4 \leq D_5)
\end{aligned}$$

Therefore, Eq. (6)-(8) do not hold for all π and all $k \in [0, j-1]$ and the algorithm will not restart.

No restarts by ENDOFREPLAYTEST. Let $\mathcal{A} \subseteq [\tau_i, t]$ be a complete replay interval of index m. Then $\Delta_{[\tau_i,t]} \leq \sqrt{\frac{KC_0}{t-\tau_i+1}} \leq \sqrt{\frac{KC_0}{|\mathcal{A}|}} = K\nu_m$. We have the following:

$$\begin{aligned}
\widehat{\text{Reg}}_{\mathcal{A}}(\pi) &\leq 2\text{Reg}_{\mathcal{A}}(\pi) + C_5\overline{K}\nu_m + C_6\Delta_{[\tau_i,t]} && \text{(Lemma 18)}\\
&\leq 2\text{Reg}_{\mathcal{B}_{j-1}}(\pi) + C_5\overline{K}\nu_m + (C_6+4)\Delta_{[\tau_i,t]} && \text{(Lemma 8)}
\end{aligned}$$

$$\leq 2\left(2\widehat{\mathrm{Reg}}_{\mathcal{B}_{j-1}}(\pi) + C_5\overline{K}\nu_{j-1} + C_6\Delta_{[\tau_i,t]}\right) + C_5\overline{K}\nu_m + (C_6+4)\Delta_{[\tau_i,t]} \qquad \text{(Lemma 16)}$$

$$\leq 4\widehat{\mathrm{Reg}}_{\mathcal{B}_{j-1}}(\pi) + D_1\overline{K}\nu_m. \qquad (3C_5 + 3C_6 + 4 \leq D_1)$$

Similarly, $\widehat{\mathrm{Reg}}_{\mathcal{B}_{j-1}}(\pi) \leq 4\widehat{\mathrm{Reg}}_{\mathcal{A}}(\pi) + D_1\overline{K}\nu_m$. Also,

$$\widehat{V}_{\mathcal{A}}(Q_m,\nu_m,\pi) \leq 6.4V_{\mathcal{A}}(Q_m,\nu_m,\pi) + \frac{80C_0}{\nu_m^2|\mathcal{A}|} \qquad (\textbf{EVENT}_1)$$

$$\leq 6.4V_{\mathcal{B}_{j-1}}(Q_m,\nu_m,\pi) + 80K + \frac{6.4\Delta_{[\tau_i,t]}}{\nu_m} \qquad \text{(Lemma 9)}$$

$$\leq 6.4\left(6.4\widehat{V}_{\mathcal{B}_{j-1}}(Q_m,\nu_m,\pi) + \frac{80C_0}{\nu_m^2|\mathcal{B}_{j-1}|}\right) + 80K + \frac{6.4\Delta_{[\tau_i,t]}}{\nu_m} \qquad (\textbf{EVENT}_1)$$

$$\leq 41\widehat{V}_{\mathcal{B}_{j-1}}(Q_m,\nu_m,\pi) + D_2K. \qquad (6.4\times 80 + 80 + 6.4 \leq D_2)$$

Therefore Eq. (3)-(5) do not hold for all π and the algorithm will not restart. ■

Proof [Lemma 7] Condition on $\textbf{EVENT}_1$, which happens with probability $1-\delta/2$. By Lemma 27, if there is no distribution change since last restart (which implies $\Delta_{[\tau_i,t]} = 0$), then the algorithm will not restart again. Thus $E \leq S$.

Let the epoch length be $T_1,\ldots,T_E$. Again by Lemma 27, in epoch i, the total variation has to be larger than $\sqrt{\frac{KC_0}{T_i}}$. By Hölder's inequality, we have

$$E-1 \leq \left(\sum_{i=1}^{E-1} T_i\right)^{\frac{1}{3}}\left(\sum_{i=1}^{E-1}\sqrt{\frac{1}{T_i}}\right)^{\frac{2}{3}} \leq T^{\frac{1}{3}}\left(\frac{\Delta}{\sqrt{KC_0}}\right)^{\frac{2}{3}} = (KC_0)^{-\frac{1}{3}}\Delta^{\frac{2}{3}}T^{\frac{1}{3}}.$$

This finishes the proof. ■

Appendix E. Notation Table

Table 1: General Notations

Notation	Meaning
$\mathcal{X}$	context space
K	number of actions
$\mathcal{D}_t$	the density function over $\mathcal{X} \times [0,1]^K$ at time t
$\mathcal{D}_t^{\mathcal{X}}$	the marginal distribution of $\mathcal{D}_t$ over the context space $\mathcal{X}$
(x_t, r_t)	context-reward pair drawn from $\mathcal{D}_t$ at time t
S	$1 + \sum_{t=2}^{T} \mathbf{1}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\}$
Δ	$\sum_{t=2}^{T} \lVert \mathcal{D}_t - \mathcal{D}_{t-1} \rVert_{\text{TV}}$
$\mathcal{I} = [s, s']$	an time interval consisting of time steps $s, s+1, \ldots, s'$
$S_{[s,s']}$	$1 + \sum_{\tau=s+1}^{s'} \mathbf{1}\{\mathcal{D}_\tau \neq \mathcal{D}_{\tau-1}\}$
$\Delta_{[s,s']}$	$\sum_{\tau=s+1}^{s'} \lVert \mathcal{D}_\tau - \mathcal{D}_{\tau-1} \rVert_{\text{TV}}$
$\mathcal{R}_t(\pi)$	$\mathbb{E}_{(x,r)\sim\mathcal{D}_t}[r(\pi(x))]$
π_t^*	$\operatorname{argmax}_{\pi\in\Pi} \mathcal{R}_t(\pi)$
$\mathcal{R}_{\mathcal{I}}(\pi)$	$\frac{1}{\lvert\mathcal{I}\rvert} \sum_{t\in\mathcal{I}} \mathcal{R}_t(\pi)$
$\pi_{\mathcal{I}}^*$	$\operatorname{argmax}_{\pi\in\Pi} \mathcal{R}_{\mathcal{I}}(\pi)$
$\text{Reg}_{\mathcal{I}}(\pi)$	$\mathcal{R}_{\mathcal{I}}(\pi_{\mathcal{I}}^*) - \mathcal{R}_{\mathcal{I}}(\pi)$
$\widehat{r}_t(a)$	$\frac{r_t(a)}{p_t(a)}\mathbf{1}\{a_t = a\}$ where $a_t \sim p_t$ is the action selected at time t
$\widehat{\mathcal{R}}_{\mathcal{I}}(\pi)$	$\frac{1}{\lvert\mathcal{I}\rvert} \sum_{t\in\mathcal{I}} \widehat{r}_t(\pi(x_t))$
$\widehat{\pi}_{\mathcal{I}}$	$\operatorname{argmax}_{\pi\in\Pi} \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)$.
$\widehat{\text{Reg}}_{\mathcal{I}}(\pi)$	$\widehat{\mathcal{R}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) - \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)$
$Q \in \Delta^{\Pi}$	$\{Q \in \mathbb{R}_+^{\lvert\Pi\rvert} : \sum_{\pi\in\Pi} Q(\pi) = 1\}$
$Q(a\lvert x)$	$\sum_{\pi:\pi(x)=a} Q(\pi)$
$Q^\nu(\cdot\lvert x)$	$\nu\mathbf{1} + (1 - K\nu)Q(\cdot\lvert x)$
$\widehat{V}_{\mathcal{I}}(Q, \nu, \pi)$	$\frac{1}{\lvert\mathcal{I}\rvert} \sum_{t\in\mathcal{I}} \left[\frac{1}{Q^\nu(\pi(x_t)\lvert x_t)}\right]$
$V_{\mathcal{I}}(Q, \nu, \pi)$	$\frac{1}{\lvert\mathcal{I}\rvert} \sum_{t\in\mathcal{I}} \mathbb{E}_{x\sim\mathcal{D}_t^{\mathcal{X}}} \left[\frac{1}{Q^\nu(\pi(x)\lvert x)}\right]$
$\overline{K}$	$(\log_2 T)K$