Efficient Contextual Bandits in Non-stationary Worlds

Haipeng Luo   HAIPENGL@USC.EDU
University of Southern California

Chen-Yu Wei   CHENYU.WEI@USC.EDU
University of Southern California

Alekh Agarwal   ALEKHA@MICROSOFT.COM
Microsoft Research, NYC

John Langford   JCL@MICROSOFT.COM
Microsoft Research, NYC

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Most contextual bandit algorithms minimize regret against the best fixed policy, a questionable benchmark for non-stationary environments that are ubiquitous in applications. In this work, we develop several efficient contextual bandit algorithms for non-stationary environments by equipping existing methods for i.i.d. problems with sophisticated statistical tests so as to dynamically adapt to a change in distribution.

We analyze various standard notions of regret suited to non-stationary environments for these algorithms, including interval regret, switching regret, and dynamic regret. When competing with the best policy at each time, one of our algorithms achieves regret $\mathcal{O}(\sqrt{ST})$ if there are T rounds with S stationary periods, or more generally $\mathcal{O}(\Delta^{1/3}T^{2/3})$ where Δ is some non-stationarity measure. These results almost match the optimal guarantees achieved by an inefficient baseline that is a variant of the classic Exp4 algorithm. The dynamic regret result is also the first one for efficient and fully adversarial contextual bandit.

Furthermore, while the results above require tuning a parameter based on the unknown quantity S or Δ, we also develop a parameter free algorithm achieving regret $\min\{S^{1/4}T^{3/4}, \Delta^{1/5}T^{4/5}\}$. This improves and generalizes the best existing result $\Delta^{0.18}T^{0.82}$ by Karnin and Anava (2016) which only holds for the two-armed bandit problem.

1. Introduction

Algorithms for the contextual bandit problem have been developed for adversarial (Auer et al., 2002), stochastic (Agarwal et al., 2014; Langford and Zhang, 2008) and hybrid (Rakhlin and Sridharan, 2016a; Syrgkanis et al., 2016b) environments. Despite the specific setting, however, almost all these works minimize the classical notion of regret that compares the reward of the algorithm to the *best fixed policy in hindsight*. This is a natural benchmark when the data generating mechanism is essentially stationary, so that a fixed policy can attain a large reward. However, in many applications of contextual bandits, we are faced with an extremely non-stationary world. For instance, the pool of available news stories or blog articles rapidly evolves in content personalization domains, and people's preferences typically exhibit trends on daily, weekly and seasonal scales. In such cases, one wants to compete with an appropriately adaptive sequence of benchmark policies, for the baseline to be meaningful.

Prior works in a context-free setting (that is, the multi-armed bandit problem) have studied regret to a sequence of actions, whenever that sequence is *slowly changing* under some appropriate measure (see e.g. (Auer et al., 2002; Besbes et al., 2014, 2015; Karnin and Anava, 2016; Wei et al., 2016)). A natural generalization to the contextual setting would be to compete with a sequence of policies, all chosen from some policy class. Extension of the prior context-free works to the contextual setting indeed yields algorithms with such guarantees, as we show with a baseline example (Exp4.S). However, the computation and storage of the resulting algorithms are both linear in the cardinality of the policy class, making tractable implementation impossible except for very small policy classes.

To overcome the computational obstacle, all previous works on efficient contextual bandits assume access to an optimization oracle which can find the policy with the largest reward on any dataset containing context-reward pairs (Langford and Zhang, 2008; Agarwal et al., 2014; Rakhlin and Sridharan, 2016a; Syrgkanis et al., 2016b). Given such an oracle, however, it is known that no efficient low-regret algorithms exist in the fully adversarial setting (Hazan and Koren, 2016, Theorem 25), even without any challenges of non-stationarity. Consequently all previous works explicitly rely on assumptions such as i.i.d. contexts, or even i.i.d. context-reward pairs.

As a warm-up and also an example to show the difficulty of the problem, we first consider a general approach to convert an algorithm for the stationary setting to an algorithm that can deal with non-stationary data. The idea is to combine different copies of the base algorithm, each of which starts at a different time to learn over different data segments. This can be seen as a natural generalization of the approach of Hazan and Seshadri (2007) for the full information setting. We build on a recent result of Agarwal et al. (2017) to deal with the additional challenges due to partial feedback and use BISTRO+ (Syrgkanis et al., 2016b) as the base algorithm since it is efficient and requires no statistical assumption on the rewards. However, unlike the full information setting, the regret rates degrade after this conversion as we show, making this general approach unsatisfying.

We next consider a more specific approach by equipping existing algorithms for the i.i.d. setting, such as EPOCH-GREEDY (Langford and Zhang, 2008) and the statistically more efficient approach of Agarwal et al. (2014), with some sophisticated statistical tests to detect non-stationarity (the resulting algorithms are called ADA-GREEDY and ADA-ILTCB respectively). Once such non-stationarity is detected, the algorithms restart from scratch. The exact tests are algorithm-specific and based on verifying certain concentration inequalities which the algorithm relies upon, but the general idea might be applicable to extending other contextual bandit algorithms as well.

We present strong theoretical guarantees for our algorithms, in terms of interval regret, switching regret and dynamic regret (defined in Section 2). A high-level outcome of our analysis is that the algorithms enjoy a regret bound on any time interval that is sufficiently stationary (called interval regret), compared with the best fixed policy for that interval. This general result has important corollaries, discussed in Section 4. For example, if the data-generating process is typically i.i.d., except there are *hard switches* in the data distribution every so often, then our algorithms perform as if they knew the change points in advance, up to a small penalty in regret (called switching regret). More generally, if the data distribution is *slowly drifting*, we can still provide meaningful regret bounds (called dynamic regret) when competing to the best policy at each time (instead of a fixed policy across all rounds).

These results are summarized in Table 1. The highlight is that our computationally efficient algorithm ADA-ILTCB enjoys almost the same guarantee as the inefficient baseline Exp4.S for all three regret measures, which is optimal in light of the existing results for the special case of

Table 1: Comparisons of different results presented in this work. "OE?" indicates whether the algorithm is Oracle-Efficient or not. T is the total number of rounds, $\mathcal{I}$ is the interval on which interval regret is measured, S is the number of i.i.d. periods, Δ is the reward variation, and $\bar{\Delta} \geq \Delta$ is the total variation, all defined in Section 2. These parameters are assumed to be known for the column "tuned" but unknown for the column "param-free".[2] Dependence on other parameters are omitted. Results for BISTRO+ assumes a transductive setting, and interval regret for the last three algorithms assumes (approximately) i.i.d. data on $\mathcal{I}$.

Algorithm	OE?	Interval Regret		Switching Regret		Dynamic Regret	
		param-free	tuned	param-free	tuned	param-free	tuned
Exp4.S (baseline)	N	$\sqrt{T}$	$\sqrt{\|\mathcal{I}\|}$	$S\sqrt{T}$	$\sqrt{ST}$	$\sqrt{\Delta}T^{\frac{2}{3}}$	$\Delta^{\frac{1}{3}}T^{\frac{2}{3}}$
Corral BISTRO+	Y	$T^{\frac{3}{4}}$	$T^{\frac{1}{4}}\sqrt{\|\mathcal{I}\|}$	$ST^{\frac{3}{4}}$	$\sqrt{S}T^{\frac{3}{4}}$	$\sqrt{\Delta}T^{\frac{5}{6}}$	$\Delta^{\frac{1}{3}}T^{\frac{5}{6}}$
ADA-GREEDY	Y	$T^{\frac{1}{6}}\sqrt{\|\mathcal{I}\|}$	$\|\mathcal{I}\|^{\frac{2}{3}}$	$\sqrt{S}T^{\frac{2}{3}}$	$S^{\frac{1}{3}}T^{\frac{2}{3}}$	$\sqrt{\Delta}T^{\frac{3}{4}}$	$\Delta^{\frac{1}{4}}T^{\frac{3}{4}}$
ADA-ILTCB	Y	$\sqrt{T}$	$\sqrt{\|\mathcal{I}\|}$	$S\sqrt{T}$	$\sqrt{ST}$	$\bar{\Delta}T^{\frac{2}{3}}$	$\bar{\Delta}^{\frac{1}{3}}T^{\frac{2}{3}}$
ADA-BINGREEDY	Y	$T^{\frac{3}{4}}$		$S^{\frac{1}{4}}T^{\frac{3}{4}}$		$\Delta^{\frac{1}{5}}T^{\frac{4}{5}}$	

multi-armed bandit. Importantly, the dynamic regret bounds for our algorithms hold under a *fully adversarial* setting.[1] As far as we know, this is the first result on adversarial and efficient contextual bandits.

All the results above, including those for Exp4.S, require tuning a parameter in terms of some unknown quantity. Otherwise the results degrade as shown in Table 1 and become vacuous when the non-stationarity measure (the number of stationary periods S or the reward variation Δ) is large. Our final contribution is a parameter-free variant of ADA-GREEDY, called ADA-BINGREEDY, which achieves better regret (even compared to Exp4.S) in the regime when S or Δ is large and unknown. Importantly, this result even improves upon the best existing result by Karnin and Anava (2016) for the context-free setting, where a regret bound of order $\Delta^{0.18}T^{0.82}$ is shown for the two-armed bandit problem. We improve the bound to $\min\{S^{1/4}T^{3/4}, \Delta^{1/5}T^{4/5}\}$ and also significantly generalize it to the multi-armed and contextual setting.

Related work. The idea of testing for non-stationarity in bandits was studied in (Bubeck and Slivkins, 2012) and (Auer and Chiang, 2016) for a very different purpose. The closest bounds to those in Table 1 are in the non-contextual setting (Auer et al., 2002; Besbes et al., 2014, 2015; Wei et al., 2016) as mentioned earlier. Chakrabarti et al. (2009) study a context-free setup where the action set changes. To the best our knowledge, oracle-efficient contextual bandit algorithms for non-stationary environments were only studied before in (Syrgkanis et al., 2016a), where a reduction from competing with a switching policy sequence to competing with a fixed policy was proposed. However, the reduction cannot be applied to the i.i.d methods (Langford and Zhang, 2008; Agarwal

---

1. Note that this does not contradict with the hardness results in (Hazan and Koren, 2016) since the bound is data-dependent and could be linear in T in the worst case.

2. Other (incomparable) bounds for the "param-free" column are also possible. See discussions in respective sections.

et al., 2014), and it heavily relies on knowing the number of switches and the transductive setting. Additionally, this approach gives no guarantees on interval regret or dynamic regret, unlike our results.

2. Preliminaries

The contextual bandits problem is defined as follows. Let $\mathcal{X}$ be an arbitrary context space and K be the number of actions. Let $[n]$ denote the set $\{1, \ldots, n\}$ for any integer n. A mapping $\pi : \mathcal{X} \to [K]$ is called a policy and the learner is given a fixed set of policies Π. For simplicity, we assume Π is a finite set but with a large cardinality $N = |\Pi|$. Ahead of time, the environment decides T distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ on $\mathcal{X} \times [0, 1]^K$, and draws T context-reward pairs $(x_t, r_t) \sim \mathcal{D}_t$ for $t = 1, \ldots, T$ independently.[3] Then at each round $t = 1, \ldots, T$, the environment reveals x_t to the learner, the learner picks an action $a_t \in [K]$ and observes its reward $r_t(a_t)$. The regret of the learner with respect to a policy π at round t is $r_t(\pi(x_t)) - r_t(a_t)$. Most existing results on contextual bandits focus on minimizing cumulative regret against any fixed policy $\pi \in \Pi$: $\sum_{t=1}^{T} r_t(\pi(x_t)) - r_t(a_t)$.

To better deal with non-stationary environments, we consider several related notions of regret. The first one is cumulative regret with respect to a fixed policy on a time interval $\mathcal{I}$, which we call *interval regret* on $\mathcal{I}$. Specifically, we use the notation $\mathcal{I} = [s, s']$ for $s \leq s'$ and $s, s' \in [T]$ to denote the set $\{s, s + 1, \ldots, s'\}$ and call it a time interval (starting from round s to round s'). The regret with respect to a fixed $\pi \in \Pi$ on a time interval $\mathcal{I}$ is then defined as $\sum_{t \in \mathcal{I}} r_t(\pi(x_t)) - r_t(a_t)$. This is similar to the notion of adaptive and strongly adaptive regret (Hazan and Seshadhri, 2007; Daniely et al., 2015). We use the term interval regret without any specific interval when the choice is clear from context.

Interval regret is useful in studying more general regret measures for non-stationary environments. Specifically, we aim at the most challenging benchmark, that is, the cumulative rewards achieved by using the best policy *at each time*. Formally, let $\mathcal{R}_t(\pi) := \mathbb{E}_{(x,r) \sim \mathcal{D}_t} r(\pi(x))$ be the expected reward of policy π under $\mathcal{D}_t$ and $\pi_t^\star := \operatorname{argmax}_{\pi \in \Pi} \mathcal{R}_t(\pi)$ be the optimal policy at round t. Then the aforementioned general regret is defined as $\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t)$. It is well-known that in general no sub-linear regret is achievable with this definition.

However, one can bound such regret in terms of some quantity that measures the non-stationarity of the environment and achieve meaningful results whenever such quantity is not too large. One example is to count the number of switches in the distribution sequence, that is, $\sum_{t=2}^{T} \mathbf{1}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\}$. We denote this by $S - 1$ (so that S is the number of i.i.d. segments) and call a regret bound in terms of S *switching regret*.

Switching regret might be meaningless if the distribution is slowly drifting, leading to a large number of switches but overall a small amount of variation in the distribution. To capture this situation, we also consider another type of non-stationarity measure, generalizing a similar notion from the multi-armed bandit literature (Besbes et al., 2014). Specifically, define $\Delta = \sum_{t=2}^{T} \max_{\pi \in \Pi} |\mathcal{R}_t(\pi) - \mathcal{R}_{t-1}(\pi)|$ to be the variation of reward distributions. Note that this is a lower bound on the sum of total variation between consecutive distributions $\bar{\Delta} = \sum_{t=2}^{T} \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}} = \sum_{t=2}^{T} \int_{[0,1]^K} \int_{\mathcal{X}} |\mathcal{D}_t(x, r) - \mathcal{D}_{t-1}(x, r)| dx dr$ (see Lemma 9 for a proof). We call regret bounds in terms of Δ or $\bar{\Delta}$ *dynamic regret*.

---

3. That is, the data generating process is oblivious to the algorithm.

All algorithms we consider construct a distribution p_t over actions at round t and then sample $a_t \sim p_t$. The importance weighted reward estimator is defined as $\widehat{r}_t(a) = \frac{r_t(a)}{p_t(a)} \mathbf{1}\{a = a_t\}$, $\forall a \in [K]$. For an interval $\mathcal{I}$, we use $\mathcal{R}_\mathcal{I}(\pi)$ and $\widehat{\mathcal{R}}_\mathcal{I}(\pi)$ to denote the average expected and empirical rewards of π over $\mathcal{I}$ respectively, that is, $\mathcal{R}_\mathcal{I}(\pi) = \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \mathcal{R}_t(\pi)$ and $\widehat{\mathcal{R}}_\mathcal{I}(\pi) = \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \widehat{r}_t(\pi(x_t))$. The empirically best policy on interval $\mathcal{I}$ is defined as $\hat{\pi}_\mathcal{I} = \operatorname{argmax}_{\pi \in \Pi} \widehat{\mathcal{R}}_\mathcal{I}(\pi)$. The number of i.i.d. periods, the reward variation, and the total variation on an interval $\mathcal{I} = [s, s']$ are respectively defined as $S_\mathcal{I} = 1 + \sum_{\tau=s+1}^{s'} \mathbf{1}\{\mathcal{D}_\tau \neq \mathcal{D}_{\tau-1}\}$, $\Delta_\mathcal{I} := \sum_{\tau=s+1}^{s'} \max_{\pi \in \Pi} |\mathcal{R}_\tau(\pi) - \mathcal{R}_{\tau-1}(\pi)|$, and $\bar{\Delta}_\mathcal{I} := \sum_{\tau=s+1}^{s'} \|\mathcal{D}_\tau - \mathcal{D}_{\tau-1}\|_{\text{TV}}$.

We use $\mathcal{D}_t^\mathcal{X}$ to denote the marginal distribution of $\mathcal{D}_t$ over $\mathcal{X}$, and $\mathbb{E}_t$ to denote the conditional expectation given everything before round t. Finally, we are interested in efficient algorithms assuming access to an optimization oracle (Agarwal et al., 2014):

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

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

In the rest of the paper, we start with discussing interval regret in Section 3, followed by the implications for switching/dynamic regret in Section 4. The parameter-free algorithm ADA-BINGREEDY is then discussed in Section 5.

3. Interval Regret

In this section we present several algorithms with interval regret guarantees. As a starter and a baseline, we first point out that a generalization of the Exp3.S algorithm (Auer et al., 2002) and Fixed-Share (Herbster and Warmuth, 1998) to the contextual bandit setting, which we call Exp4.S, already provides a strong interval regret guarantee as shown by the following theorem. We include the algorithm and the proof in Appendix B. Crucially, Exp4.S requires maintaining weights for each policy and is thus not oralce-efficient.

Theorem 2 *Exp4.S with parameter L ensures that for any time interval $\mathcal{I}$ such that $|\mathcal{I}| \leq L$, we have $\mathbb{E}\left[\sum_{t \in \mathcal{I}} r_t(\pi(x_t)) - r_t(a_t)\right] \leq \mathcal{O}(\sqrt{LK \ln(NL)})$ for any $\pi \in \Pi$, where the expectation is with respect to the randomness of both the algorithm and the environment.*

Note that in bandit settings, it is impossible to achieve regret $\mathcal{O}(\sqrt{|\mathcal{I}|})$ for all interval $\mathcal{I}$ simultaneously (Daniely et al., 2015). When $|\mathcal{I}|$ is unknown, a safe choice is to pick $L = T$ (this is how we obtain the results in the "param-free" column of Table 1 for interval regret). Next we prove statements similar to Theorem 2 but with oracle-efficient algorithms.

A general approach. In the full information setting, a general approach to convert an algorithm with classic regret guarantee to another with interval regret is to combine different copies of the algorithm with an expert algorithm, each of which starts at a different time step to learn over different time intervals. This works well in the full information setting where one has correct feedback to update all the base algorithms, but becomes challenging in the bandit setting. We show in Appendix G how to leverage recent results by Agarwal et al. (2017) and Wei and Luo (2018) to deal with such challenges. As an example we use the BISTRO+ algorithm (Syrgkanis et al., 2016b; Rakhlin and Sridharan, 2016b) as the base algorithm since it is oracle-efficient and allows adversarial rewards.

Theorem 3 *In the transductive setting, Algorithm 5 in Appendix G guarantees that for any time interval $\mathcal{I}$ such that $|\mathcal{I}| \leq L$ and any policy $\pi \in \Pi$, we have $\mathbb{E}\left[\sum_{t \in \mathcal{I}} r_t(\pi(x_t)) - r_t(a_t)\right] \leq \widetilde{\mathcal{O}}(T^{\frac{1}{4}}(LK)^{\frac{1}{2}}(\ln N)^{\frac{1}{4}}).$*

Unlike the full information setting, this general approach results in worse regret rates and is unsatisfying (BISTRO+ achieves $\widetilde{\mathcal{O}}(T^{2/3})$ for the classic regret and here we only obtain $\widetilde{\mathcal{O}}(T^{3/4})$). In the following subsections, we turn to different approaches.

3.1. ADA-GREEDY

The simplest oracle-efficient contextual bandit algorithm is the EPOCH-GREEDY method (Langford and Zhang, 2008) which assumes i.i.d. data. In this section, we extend the related ϵ-GREEDY algorithm to enjoy a small interval regret on any interval with a small variation.

ϵ-GREEDY plays uniformly at random with a small probability and otherwise follows the empirically best policy $\pi_t = \arg\max_{\pi \in \Pi} \widehat{\mathcal{R}}_{[1,t-1]}(\pi)$. The number of oracle calls can be greatly reduced if the learner updates the best policy only at $t = 1, 2, 4, 8, \ldots$ (that is, $\pi_t = \arg\max_{\pi \in \Pi} \widehat{\mathcal{R}}_{[1,2^{\lfloor \log_2 t \rfloor} - 1]}(\pi)$). ADA-GREEDY, described in Algorithm 1, behaves similarly to this version of ϵ-GREEDY. The difference is that at each round, an additional *non-stationarity test* is executed. The test monitors whether there is a policy performing significantly better on recent samples (collected in a doubling manner), compared to the policy that the algorithm is using. Intuitively, if such a policy exists, there should have been a significant shift in the distribution. In this case, the algorithm restarts from scratch.

In addition, the algorithm also resets every L rounds for some parameter L (Line 9). This prevents the risk of slow detection of a distribution change, but at the same time also causes some extra penalty when the environment is stationary. The parameter L trades these two kinds of costs, and can be selected based on prior knowledge about the environment.

We call the rounds between resets an *epoch* (so epoch i is the interval $[T_i + 1, T_{i+1}]$), and the rounds between updates of the empirically best policy a *block* (so block j of epoch i is the interval $[T_i + 2^{j-1}, T_i + 2^j - 1]$).

Note that there are only two places where we need to invoke the oracle: computing $\hat{\pi}_{(i,j)}$ and $\hat{\pi}_A$ ($\hat{\pi}_B$ is simply equal to $\hat{\pi}_{(i,j)}$), and it is thus clear that at most $\mathcal{O}(\ln L) = \mathcal{O}(\ln T)$ oracle calls are used per round.

We prove the following result for ADA-GREEDY, stating a regret bound for all intervals with length smaller than L and variation smaller than another parameter v of the algorithm.

Theorem 4 *With probability at least $1 - \delta$, for all time intervals $\mathcal{I}$ such that $|\mathcal{I}| \leq L$ and $\Delta_{\mathcal{I}} \leq v$, ADA-GREEDY with parameters L, v and δ guarantees for any $\pi \in \Pi$,[4]*

$$\sum_{t \in \mathcal{I}} r_t(\pi(x_t)) - r_t(a_t) \leq \widetilde{\mathcal{O}}\left(|\mathcal{I}|v + L^{\frac{1}{6}}\sqrt{K|\mathcal{I}|\ln(N/\delta)} + K\ln(N/\delta)\right).$$

Note that whenever $v = O(L^{-\frac{1}{3}})$, the rate of the regret above is of order $\widetilde{\mathcal{O}}(L^{2/3})$ (since $|\mathcal{I}| \leq L$), which matches the ordinary regret bound of EPOCH-GREEDY ($\widetilde{\mathcal{O}}(T^{2/3})$). While a condition on both the interval length and variation is seemingly strong and the bound seems to be meaningful only for very small v, we emphasize that 1) sublinear regret via oracle-efficient algorithms is impossible

---

4. We use notation $\widetilde{\mathcal{O}}$ to suppress dependence on logarithmic factors in L, T, K and $\ln(N/\delta)$.

Algorithm 1 ADA-GREEDY

1 **Input**: largest allowed interval length L and variation v, allowed failure probability δ

2 **Define**: $\mu = \min\left\{\frac{1}{K}, L^{-\frac{1}{3}}\sqrt{\frac{\ln(N/\delta)}{K}}\right\}, \beta_{\mathcal{I}} = 2\sqrt{\frac{\ln(4T^2N/\delta)}{\mu|\mathcal{I}|}} + \frac{\ln(4T^2N/\delta)}{\mu|\mathcal{I}|}$

3 **Initialize**: $i = 1, T_1 = 0$.   ▷ i indexes an epoch

4 **for** $j = 1, 2, \ldots$ **do**   ▷ j indexes a block

5 $\quad$ Compute $\hat{\pi}_{(i,j)} = \mathrm{argmax}_{\pi \in \Pi} \widehat{\mathcal{R}}_{[T_i+1, T_i+2^{j-1}-1]}(\pi)$ ▷ or arbitrary if $j = 1$

6 $\quad$ **for** $t = T_i + 2^{j-1}, \ldots, T_i + 2^j - 1$ **do**

7 $\quad\quad$ Set $p_t(a) = \mu + (1 - K\mu)\mathbf{1}\{a = \hat{\pi}_{(i,j)}(x_t)\}, \forall a \in [K]$

8 $\quad\quad$ Play $a_t \sim p_t$ and receive $r_t(a_t)$

9 $\quad\quad$ **if** $(t \geq T_i + L)$ *or* $(j > 1$ *and* NONSTATTEST$(t) = $ *True*$)$ **then**

10 $\quad\quad\quad$ $T_{i+1} \leftarrow t, i \leftarrow i + 1$

11 $\quad\quad\quad$ **goto** Line 4

$\quad$

Procedure NONSTATTEST(t)

12 $\ell = 1$

13 **while** $\ell \leq t - T_i$ **do**

14 $\quad$ Let $A \triangleq [t - \ell + 1, t]$ and $B \triangleq [T_i + 1, T_i + 2^{j-1} - 1]$

15 $\quad$ **if** $\widehat{\mathcal{R}}_A(\hat{\pi}_A) > \widehat{\mathcal{R}}_A(\hat{\pi}_B) + 2(\beta_A + \beta_B + 2v)$ **then** **return** *True*

16 $\quad$ $\ell \leftarrow 2\ell$

17 **return** *False*

---

under a fully adversarial setting even for the classic regret (Hazan and Koren, 2016) and 2) based on Theorem 4 we can in fact derive strong dynamic regret bounds that hold without any assumption on the distribution sequence (see Section 4).

3.2. ADA-ILTCB

Although being fairly simple, ADA-GREEDY is suboptimal just as EPOCH-GREEDY is suboptimal for stationary environments. In this section we propose ADA-ILTCB, a variant of ILOVETO-CONBANDITS (Agarwal et al., 2014), which achieves the optimal regret rate while also being oracle-efficient. The idea is similar to ADA-GREEDY, but the statistical checks are more involved.

For a policy π and an interval $\mathcal{I}$, we denote the expected and empirical regret of π by $\mathrm{Reg}_{\mathcal{I}}(\pi) = \max_{\pi' \in \Pi} \mathcal{R}_{\mathcal{I}}(\pi') - \mathcal{R}_{\mathcal{I}}(\pi)$ and $\widehat{\mathrm{Reg}}_{\mathcal{I}}(\pi) = \max_{\pi' \in \Pi} \widehat{\mathcal{R}}_{\mathcal{I}}(\pi') - \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)$ respectively. For a context x and a distribution over the policies $Q \in \Delta^{\Pi} := \{Q \in \mathbb{R}_+^N : \sum_{\pi \in \Pi} Q(\pi) = 1\}$, the projected distribution over the actions is denoted by $Q(\cdot|x)$ such that $Q(a|x) = \sum_{\pi:\pi(x)=a} Q(\pi), \ \forall a \in [K]$. The smoothed projected distribution with a minimum probability μ is defined as $Q^{\mu}(\cdot|x) = \mu\mathbf{1} + (1 - K\mu)Q(\cdot|x)$ where $\mathbf{1}$ is the all-one vector. Like (Agarwal et al., 2014), we keep track of a bound on the variance of the reward estimates and define for a policy π, an interval $\mathcal{I}$ and a distribution $Q \in \Delta^{\Pi}$

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

Algorithm 2 ADA-ILTCB

---

1 **Input**: largest allowed interval length L and variation v, allowed failure probability δ

2 **Define**: $\mu = \min\left\{\frac{1}{2K}, L^{-\frac{1}{2}}\sqrt{\frac{\ln(8T^2N^2/\delta)\ln(L)}{K}}\right\}, C_1 = 4, C_2 = 10^6, C_3 = 1.1 \times 10^3, C_4 = 41, C_5 = 1200, C_6 = 6.4$

3 **Initialize**: $i = 1, T_1 = 0$ ▷ i indexes an epoch

4 **for** $j = 1, 2, \ldots$ **do**   ▷ j indexes a block

5   Let $Q_{(i,j)}$ be a solution to (OP) with parameter μ and data from $[T_i + 1, T_i + 2^{j-1} - 1]$

6   **for** $t = T_i + 2^{j-1}, \ldots, T_i + 2^j - 1$ **do**

7   Set $p_t(a) = Q_t^\mu(a|x_t), \forall a \in [K]$ where $Q_t = Q_{(i,j)}$

8   Play $a_t \sim p_t$ and receive $r_t(a_t)$

9   **if** $(t \geq T_i + L)$ *or* (NONSTATTEST$(t) = $ *True*) **then**

10              $T_{i+1} \leftarrow t, i \leftarrow i + 1$

11   **goto** Line 4

Procedure NONSTATTEST(t)

12 $\ell = 1$

13 **while** $\ell \leq t - T_i - 1$ **do**

14   Let $A \triangleq [t - \ell, t - 1]$ and $B \triangleq [T_i + 1, T_i + 2^{j-1} - 1]$

15   **if** $\max_{\pi \in \Pi}\left\{\widehat{\mathrm{Reg}}_B(\pi) - C_1\widehat{\mathrm{Reg}}_A(\pi)\right\} > \frac{C_2 LK\mu}{\ell} + C_3 v$ **then** **return** *True*

16   **if** $\max_{\pi \in \Pi}\left\{\widehat{\mathrm{Reg}}_A(\pi) - C_1\widehat{\mathrm{Reg}}_B(\pi)\right\} > \frac{C_2 LK\mu}{\ell} + C_3 v$ **then** **return** *True*

17   **if** $\max_{\pi \in \Pi}\left\{\widehat{V}_A(Q_t, \pi) - C_4\widehat{V}_B(Q_t, \pi)\right\} > \frac{C_5 LK}{\ell} + \frac{C_6 v}{\mu}$ **then** **return** *True*

18      $\ell \leftarrow 2\ell$

19 **return** *False*

Similar to ADA-GREEDY, the proposed algorithm ADA-ILTCB (Algorithm 2) proceeds like the base algorithm (ILOVETOCONBANDITS in this case) with additional tests to detect the non-stationarity of the environment. We define an epoch and a block similar to those of ADA-GREEDY. The algorithm solves the optimization (OP) defined in (Agarwal et al., 2014) (and included in Appendix D) at the beginning of each block using the data collected in that epoch so far. The solution of (OP) is denoted by Q_t, a *sparse* distribution over Π, and the learner samples actions based on $Q_t^\mu(\cdot|x_t)$.

At each round, the NONSTATTEST checks whether the empirical regret or the variance of reward estimates of any policy has changed significantly in a recent interval (i.e., $[t - \ell, t - 1]$), compared to the interval from which we compute Q_t (i.e., $[T_i + 1, T_i + 2^{j-1} - 1]$). If so, the algorithm restarts with a new epoch. Note that detecting the change of regret is similar to detecting the change of reward; but different from ADA-GREEDY, here we also check the change of reward estimate variance. This inherits from the tighter variance control in ILOVETOCONBANDITS, the key to obtaining better regret compared to ϵ-GREEDY.

Oracle-Efficiency. Note that Lines 15, 16 and 17 can all be implemented by one call of the AMO oracle each, after using two extra oracle calls to compute $\max_{\pi' \in \Pi}\widehat{\mathcal{R}}_B(\pi')$ and $\max_{\pi' \in \Pi}\widehat{\mathcal{R}}_A(\pi')$ in advance. Specifically, let $\mathcal{S} = \{(x_\tau, \frac{-1}{2^{j-1}-1}\widehat{r}_\tau)\}_{\tau \in B} \cup \{(x_\tau, \frac{C_1}{\ell}\widehat{r}_\tau)\}_{\tau \in A}$, then the left hand

side of the inequality in Line 15 can be rewritten as $\max_\pi \sum_{(x,r)\in\mathcal{S}} r(\pi(x)) + \max_{\pi'\in\Pi} \widehat{\mathcal{R}}_B(\pi') - C_1 \max_{\pi'\in\Pi} \widehat{\mathcal{R}}_A(\pi')$, where clearly the first term can be computed by one oracle call and the rests are precomputed already. Similarly, Line 17 can be computed by feeding the oracle with examples $\{(x_\tau, \frac{1}{\ell}\frac{1}{Q_\tau^\mu(\cdot|x_\tau)})\}_{\tau\in A} \cup \{(x_\tau, \frac{-C_4}{2^{j-1}-1}\frac{1}{Q_\tau^\mu(\cdot|x_\tau)})\}_{\tau\in B}$.

Agarwal et al. (2014) showed that the optimization problem (OP) can be solved by $\widetilde{\mathcal{O}}(1/\mu)$ oracle calls and the solution has only $\widetilde{\mathcal{O}}(1/\mu)$ non-zero coordinates. Note that we only solve (OP) at the beginning of each block. Since there are $\mathcal{O}(\ln L)$ blocks in an epoch, the total oracle calls in an epoch is bounded by $\widetilde{\mathcal{O}}(\ln(L)/\mu) = \widetilde{\mathcal{O}}(\sqrt{LK})$, which amortizes to $\widetilde{\mathcal{O}}(S'\sqrt{LK}/T)$ per round if there are S' epochs (in Section 4 we relate S' to S or $\bar{\Delta}$).

We next present the interval regret guarantee of ADA-ILTCB, which improves from $\widetilde{\mathcal{O}}(L^{\frac{2}{3}})$ to $\widetilde{\mathcal{O}}(\sqrt{L})$ compared to ADA-GREEDY (see Appendix D for the proof), except that it holds for interval with total variation $\bar{\Delta}_\mathcal{I}$ (instead of reward variation $\Delta_\mathcal{I}$) bounded by v due to the fact that variation in the context is important for the variance control (Line 17).

Theorem 5 *With probability at least $1 - \delta$, for any interval $\mathcal{I}$ such that $|\mathcal{I}| \leq L$ and $\bar{\Delta}_\mathcal{I} \leq v$,* ADA-ILTCB *with parameters L, v, and δ guarantees for any $\pi \in \Pi$,*

$$\sum_{t\in\mathcal{I}} r_t(\pi(x_t)) - r_t(a_t) \leq \widetilde{\mathcal{O}}\left(|\mathcal{I}|v + \sqrt{LK\ln(N/\delta)}\right).$$

4. Implications

In this section we discuss the implications of interval regret guarantees on switching/dynamic regret, both of which are meaningful performance measures for non-stationary environments.

Switching Regret. We begin with switching regret, which is pretty straightforward. One only needs to divide the entire time interval $[1, T]$ into several i.i.d. subintervals with length bounded by L, and then apply the interval regret guarantee on each of these subintervals since the best policy $\pi_t^\star$ remains the same on each of these subintervals (for ADA-GREEDY and ADA-ILTCB we can simply set the variation tolerance v to be 0). We take Exp4.S as an example and state the results below (see Appendix B for the proof), while similar results for other algorithms are summarized in Table 1.

Corollary 1 *Exp4.S with parameter L ensures* $\mathbb{E}\left[\sum_{t=1}^T r_t(\pi_t^\star(x_t)) - r_t(a_t)\right] \leq \widetilde{\mathcal{O}}\left(\left(\frac{T}{\sqrt{L}} + S\sqrt{L}\right)\sqrt{K\ln N}\right)$ *where $S = 1 + \sum_{t=2}^T \mathbf{1}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\}$.*

If S is known, then setting $L = T/S$ gives a bound of $\widetilde{\mathcal{O}}(\sqrt{STK\ln N})$. Otherwise setting L with different values leads to different bounds that are incomparable. For example, setting $L = T$ leads to $\widetilde{\mathcal{O}}(S\sqrt{TK\ln N})$ while setting $L = \sqrt{T}$ leads to $\widetilde{\mathcal{O}}((T^{\frac{3}{4}} + ST^{\frac{1}{4}})\sqrt{K\ln N})$. No matter how L is tuned, however, these bounds all become vacuous ($\Omega(T)$) when S is large enough but still sublinear in T, an issue addressed later in Section 5.

Dynamic Regret. We now move on to discuss dynamic regret in terms of the variation measures Δ or $\bar{\Delta}$ (recall $\Delta = \sum_{t=2}^T \max_{\pi\in\Pi} |\mathcal{R}_t(\pi) - \mathcal{R}_{t-1}(\pi)|$ and $\bar{\Delta} = \sum_{t=2}^T \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}}$). We first point out that previous works (Besbes et al., 2015; Zhang et al., 2017) have studied a reduction from dynamic regret to interval regret, restated below:

9

Lemma 6 *Let $\{\mathcal{I}_i = [s_i, t_i]\}_{i \in [n]}$ be time intervals that partition $[1, T]$. We have*

$$\sum_{t=1}^{T} \mathbb{E}_t \left[r_t(\pi_t^\star(x_t)) - r_t(a_t) \right] \leq \sum_{i=1}^{n} \sum_{t \in \mathcal{I}_i} \mathbb{E}_t \left[\left(r_t(\pi_{s_i}^\star(x_t)) - r_t(a_t) \right) \right] + 2 \sum_{i \in [n]} |\mathcal{I}_i| \Delta_{\mathcal{I}_i}.$$

We include the proof in Appendix F for completeness. Partitioning $[1, T]$ into intervals with equal length $L' \leq L$, applying this lemma and Theorem 2, and using the fact $\sum_{i \in [n]} \Delta_{\mathcal{I}_i} \leq \Delta$ directly lead to the following result for Exp4.S.

Corollary 2 *Exp4.S with parameter L ensures that $\mathbb{E}\left[\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t) \right] \leq \widetilde{\mathcal{O}}\left(\min_{0 \leq L' \leq L} \left\{ \frac{T}{L'} \sqrt{LK \ln N} + L' \Delta \right\} \right).$*

Again, if Δ is known one can tune L optimally to get a bound $\widetilde{\mathcal{O}}(T^{\frac{2}{3}}(\Delta K \ln N)^{\frac{1}{3}} + \sqrt{TK \ln N})$, similar to the optimal dynamic regret in multi-armed bandits (Besbes et al., 2014). When Δ is unknown, different values of L give different and in general incomparable bounds. For example, setting $L = T^{\frac{2}{3}}$ leads to $\widetilde{\mathcal{O}}(T^{\frac{2}{3}} \sqrt{\Delta}(K \ln N)^{\frac{1}{4}} + T^{\frac{2}{3}} \sqrt{K \ln N})$ (with $L' = \min\{T^{2/3}, T^{2/3}(K \ln N)^{1/4}/\sqrt{\Delta}\}$ in this case), which is again vacuous for large Δ.

Similar arguments also provide a dynamic regret bound for CORRAL with BISTRO+ in the transductive setting, as shown in Table 1 (also see Corollary 5 in Appendix G). However, the exact same argument above does not apply to ADA-GREEDY and ADA-ILTCB directly since its interval regret guarantee requires $\Delta_{\mathcal{I}} \leq v$. It turns out, however, one can set v to some carefully selected value and partition $[1, T]$ correspondingly so that every subinterval satisfies $|\mathcal{I}| \leq L$ and $\Delta_{\mathcal{I}} \leq v$, to obtain the following results that hold in a completely adversarial setting.[5] The proofs are included in Appendix F.

Corollary 3 *With probability at least $1 - \delta$, ADA-GREEDY with parameter L, δ and $v = L^{-1/3}$ ensures that*

$$\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t) \leq \widetilde{\mathcal{O}}\left(\left(\frac{T}{L^{1/3}} + L^{1/3} \sqrt{\Delta T} \right) K \ln(N/\delta) \right).$$

Specifically, if Δ is known, setting $L = \min\{(T/\Delta)^{\frac{3}{4}}, T\}$ gives $\widetilde{\mathcal{O}}((\Delta^{\frac{1}{4}} T^{\frac{3}{4}} + T^{\frac{2}{3}}) K \ln(N/\delta))$; otherwise, setting $L = T^{\frac{3}{4}}$ gives $\widetilde{\mathcal{O}}((\sqrt{\Delta} + 1) T^{\frac{3}{4}} K \ln(N/\delta))$.

Corollary 4 *With probability at least $1 - \delta$, ADA-ILTCB with parameter L, δ and $v = L^{-\frac{1}{2}}$ ensures that*

$$\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t) \leq \widetilde{\mathcal{O}}\left(\left(\frac{T}{\sqrt{L}} + \bar{\Delta} L \right) K \ln(N/\delta) \right).$$

If $\bar{\Delta}$ is known, setting $L = \min\{(T/\bar{\Delta})^{\frac{2}{3}}, T\}$ gives $\widetilde{\mathcal{O}}((\bar{\Delta}^{\frac{1}{3}} T^{\frac{2}{3}} + \sqrt{T}) K \ln(N/\delta))$; otherwise, setting $L = T^{\frac{2}{3}}$ gives $\widetilde{\mathcal{O}}((\bar{\Delta} + 1) T^{\frac{2}{3}} K \ln(N/\delta))$.

One can see that again the result for ADA-ILTCB is better than that of ADA-GREEDY, and is in fact very close to that of the inefficient baseline Exp4.S, except that it is in terms of the slightly larger variation measure $\bar{\Delta}$.

---

5. The dependence on $K \ln(N/\delta)$ in these results is slightly loose for conciseness and could be tightened.

Algorithm 3 ADA-BINGREEDY

1 **Input**: allowed failure probability δ

2 **Define**: $\beta_{\mathcal{I}} = 2\sqrt{\frac{\ln(4T^2N/\delta)}{\mu_{\mathcal{I}}|\mathcal{I}|}} + \frac{\ln(4T^2N/\delta)}{\mu_{\mathcal{I}}|\mathcal{I}|}$, where $\mu_{\mathcal{I}} \triangleq \min\limits_{t\in\mathcal{I}} \mu_t$ and μ_t is defined below, $\alpha_{\mathcal{I}} = 2\sqrt{\frac{K\ln(4T^2N/\delta)}{|\mathcal{I}|}} + \frac{K\ln(4T^2N/\delta)}{|\mathcal{I}|}$

3 **Initialize**: $t = 1, i = 1, T_1 = 0$ ▷ i indexes an epoch

4 **for** $j = 1, 2, \ldots$ **do**   ▷ j indexes a block

5   Compute $\hat{\pi}_{(i,j)} = \text{argmax}_{\pi\in\Pi} \widehat{\mathcal{R}}_{[T_i+1, T_i+2^{j-1}-1]}(\pi)$ ▷ or arbitrary if $j = 1$

6       $H = 2^{j-1}$ ▷ H is block length

7   **for** $b = 1, 2, \ldots, \sqrt{H}$ **do**   ▷ b indexes a bin, each with length $\sqrt{H}$

8   Make bin b an exploration bin with probability $1/\sqrt{b}$; otherwise an exploitation bin

9   **for** $\tau = 1, \ldots, \sqrt{H}$ **do**   ▷ loop through rounds in bin b

10   Let $\mu_t = \min\left\{\frac{1}{K}, (t-T_i)^{-\frac{1}{3}}\sqrt{\frac{\ln(N/\delta)}{K}}\right\}$

11   Set $p_t(a) = \begin{cases} \frac{1}{K}, & \text{if bin } b \text{ is an exploration bin,} \\ \mu_t + (1-K\mu_t)\mathbf{1}\{a = \hat{\pi}_{(i,j)}(x_t)\}, & \text{if bin } b \text{ is an exploitation bin.} \end{cases}$

12   Play $a_t \sim p_t$ and receive $r_t(a_t)$

13   **if** $j > 1$ *and (bin b is exploration bin) and (*NONSTATTEST$(t) = $ *True)* **then**

14                 $T_{i+1} \leftarrow t, t \leftarrow t+1, i \leftarrow i+1$

15   **goto** Line 4

16             $t \leftarrow t+1$

Procedure NONSTATTEST(t)

17 $\ell = 1$

18 **while** $[t-\ell+1, t]$ *is a subset of the current bin* **do**

19   Let $A \triangleq [t-\ell+1, t]$ and $B \triangleq [T_i+1, T_i+2^{j-1}-1]$

20   **if** $\widehat{\mathcal{R}}_A(\hat{\pi}_A) > \widehat{\mathcal{R}}_A(\hat{\pi}_B) + 2(\alpha_A + \beta_B)$ **then return** *True*

21       $\ell \leftarrow 2\ell$

22 **return** *False*

---

5. Achieving Switching/Dynamic Regret with No Parameters

As mentioned, when the parameter S or Δ is unknown, our algorithms achieve regret of the form $\widetilde{\mathcal{O}}(S^{c_1}T^{c_2})$ or $\widetilde{\mathcal{O}}(\Delta^{c_1}T^{c_2})$ for some exponents c_1 and c_2 such that $c_1 + c_2 > 1$, which is vacuous when S or Δ is large. The hope here is to obtain a bound with $c_1 + c_2 = 1$ as in the case when the parameters are known. Observe that if an algorithm was able to achieve interval regret $o(|\mathcal{I}|)$ simultaneously for all intervals $\mathcal{I}$, which is called strongly adaptive algorithm (Daniely et al., 2015), then by similar reductions discussed in Section 4 one could derive switching/dynamic regret with $c_1 + c_2 = 1$. However, it was shown by Daniely et al. (2015) that a strongly adaptive algorithm is impossible for the bandit setting.

Despite this negative result, Karnin and Anava (2016) developed new techniques and proposed a parameter-free algorithm for the two-armed bandit setting with dynamic regret $\widetilde{\mathcal{O}}(\Delta^{0.18}T^{0.82})$.

While their algorithm and analysis do not directly generalize to the multi-armed or contextual setting, here we extract their idea of *bin-based exploration* and incorporate it into our ADA-GREEDY algorithm, leading to a parameter-free algorithm called ADA-BINGREEDY with regret $\widetilde{\mathcal{O}}(\min\{S^{\frac{1}{4}}T^{\frac{3}{4}}, \Delta^{\frac{1}{5}}T^{\frac{4}{5}}\})$. This improves and generalizes the result of Karnin and Anava (2016) significantly.

Similar to ADA-GREEDY, ADA-BINGREEDY computes the empirical best policy at the beginning of each block, and plays it throughout that block, except for some exploration steps. The differences are 1) each block is further divided into *bins* with equal length; 2) in addition to the small probability of exploration μ_t at each round, some bins are randomly selected for pure exploration; 3) the non-stationarity test is only executed in exploration bins, and only checks for intervals within the bin; 4) parameters L and v are removed and the exploration probability μ_t is set adaptively. Clearly ADA-BINGREEDY is still oracle-efficient.

Comparing the non-stationarity tests of ADA-GREEDY and ADA-BINGREEDY, one can see that the term $\beta_A = \widetilde{\mathcal{O}}(1/\sqrt{\mu\ell})$ in the former is replaced by the term $\alpha_A = \widetilde{\mathcal{O}}(\sqrt{K/\ell})$ in the latter. This is due to the lower variance of reward estimates from the pure exploration bin and plays a crucial role in our analysis to achieve the following bound.

Theorem 7 *With probability at least $1 - 6\delta$,* ADA-BINGREEDY *with parameter δ guarantees*

$$\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t) \leq \widetilde{\mathcal{O}}\left(K\ln(N/\delta)\min\left\{S^{\frac{1}{4}}T^{\frac{3}{4}}, \Delta^{\frac{1}{5}}T^{\frac{4}{5}} + T^{\frac{3}{4}}\right\}\right).$$

This bound is sublinear as long as S or Δ is sublinear, and is stronger than those in the "param-free" column of Table 1 if $S = \Omega(T^{1/3})$ or $\Delta = \Omega(T^{4/9})$ (but still sublinear). One might wonder whether combining the bin-based exploration idea with ADA-ILTCB leads to even better results. The answer is unfortunately no because the dominant part of the regret is not from the ϵ-greedy part of the algorithm but the bin explorations. We leave the question of whether better results of this kind are possible as a future direction.

Due to the existence of exploration bins, ADA-BINGREEDY can have poor regret on some intervals. In fact, we can only show a loose $\widetilde{\mathcal{O}}(T^{\frac{3}{4}})$ interval regret bound for this algorithm as shown in Table 1. For completeness, we provide a proof in Appendix H.

6. Conclusions

In this work we take the first step in studying the problem of non-stationary contextual bandit. We propose several new algorithms and provide a number of achievable results under various regret notions. More future directions include 1) deriving algorithms with long term memory so as to identify distributions experienced before (Bousquet and Warmuth, 2002); 2) designing simpler and more practical algorithms, given that our current methods have several impractical aspects such as restarting.

References

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

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *30th Annual Conference on Learning Theory (COLT)*, 2017.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory (COLT)*, 2016.

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.

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.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *25th Annual Conference on Learning Theory (COLT)*, 2012.

Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. In *Advances in neural information processing systems*, pages 273–280, 2009.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

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 C. Seshadhri. Adaptive algorithms for online decision problems. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 14, 2007.

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

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*, 2008.

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

Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1977–1985, 2016b.

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*, 2016b.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. *arXiv preprint arXiv:1801.03265*, 2018.

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.

Lijun Zhang, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Strongly adaptive regret implies optimally dynamic regret. *arXiv preprint arXiv:1701.07570*, 2017.

Appendix A. Preliminaries

Our analysis relies on the following Freedman's inequality.

Lemma 8 ((Beygelzimer et al., 2011)) *Let $X_1, \ldots, X_n \in \mathbb{R}$ be a sequence of random variables such that $X_i \leq R$ and $\mathbb{E}[X_i|X_{i-1}, \ldots, X_1] = 0$ for all $i \in [n]$. Then for any $\delta \in (0,1)$ and $\lambda \in [0, 1/R]$, with probability at least $1 - \delta$, we have*

$$\sum_{i=1}^{n} X_i \leq (e-2)\lambda V + \frac{\ln(1/\delta)}{\lambda}$$

where $V = \sum_{i=1}^{n} \mathbb{E}[X_i^2|X_{i-1}, \ldots, X_1]$. Specifically, picking $\lambda = \min\left\{\sqrt{\frac{\ln(1/\delta)}{V}}, \frac{1}{R}\right\}$, we have $\sum_{i=1}^{n} X_i \leq 2\sqrt{V\ln(1/\delta)} + R\ln(1/\delta)$.

The following lemmas relates the two variation notions we use.

Lemma 9 *For any interval $\mathcal{I}$, $\Delta_{\mathcal{I}} \leq \bar{\Delta}_{\mathcal{I}}$.*

Proof Let π be any policy,

$$
\begin{aligned}
\left|\mathcal{R}_t(\pi) - \mathcal{R}_{t-1}(\pi)\right| &= \left|\mathbb{E}_{(x,r)\sim\mathcal{D}_t}[r(\pi(x))] - \mathbb{E}_{(x,r)\sim\mathcal{D}_{t-1}}[r(\pi(x))]\right| \\
&= \left|\int_{[0,1]^K}\int_{\mathcal{X}}(\mathcal{D}_t(x,r) - \mathcal{D}_{t-1}(x,r))r(\pi(x))dxdr\right| \\
&\leq \int_{[0,1]^K}\int_{\mathcal{X}}\left|\mathcal{D}_t(x,r) - \mathcal{D}_{t-1}(x,r)\right|dxdr \\
&= \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\mathrm{TV}}.
\end{aligned}
$$

Thus, $\max_{\pi\in\Pi}|\mathcal{R}_t(\pi) - \mathcal{R}_{t-1}(\pi)| \leq \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\mathrm{TV}}$. Summing over $\mathcal{I}$ gives $\Delta_{\mathcal{I}} \leq \bar{\Delta}_{\mathcal{I}}$. ■

Appendix B. Exp4.S Algorithm and Proofs

Algorithm 4 Exp4.S

Input: largest interval length of interest L

Define $\eta = \sqrt{\ln(NL)/LK}$ and $\mu = 1/NL$

Initialize $P_t \in \Delta^\Pi$ to be the uniform distribution over policies.

for $t = 1, \ldots, T$ **do**

    see x_t, play $a_t \sim p_t$ where $p_t(a) = \sum_{\pi:\pi(x_t)=a} P_t(\pi)$, $\forall a \in [K]$

    receive $r_t(a_t)$ and construct $\widehat{c}_t(a) = \frac{1-r_t(a)}{p_t(a)}\mathbf{1}\{a = a_t\}$, $\forall a \in [K]$

    set $\tilde{P}_{t+1}(\pi) \propto P_t(\pi)\exp(-\eta\widehat{c}_t(\pi(x_t)))$, $\forall \pi \in \Pi$

    set $P_{t+1}(\pi) = (1-N\mu)\tilde{P}_{t+1}(\pi) + \mu$, $\forall \pi \in \Pi$

end

The Exp4.S algorithm is presented in Algorithm 4, which is a direct generalization of Exp3.S (Auer et al., 2002). Note that we use loss estimates $\widehat{c}_t$ instead of reward estimate $\widehat{r}_t$ in the multiplicative update, and naturally we define $c_t = \mathbf{1} - r_t$.

Proof [Proof of Theorem 2] Using the fact $e^{-y} \leq 1 - y + y^2$ for any $y \geq 0$, $\ln(1 + y) \leq y$ and $c_t(a) \in [0, 1]$, we have

$$\ln\left(\sum_{\pi' \in \Pi} P_t(\pi') \exp(-\eta\widehat{c}_t(\pi'(x_t)))\right) \leq \ln\left(\sum_{\pi' \in \Pi} P_t(\pi')(1 - \eta\widehat{c}_t(\pi'(x_t)) + \eta^2\widehat{c}_t(\pi'(x_t))^2)\right)$$
$$= \ln\left(1 - \eta c_t(a_t) + \eta^2\widehat{c}_t(a_t)c_t(a_t)\right) \leq -\eta c_t(a_t) + \eta^2\widehat{c}_t(a_t).$$

On the other hand, we have for any fixed π,

$$\ln\left(\sum_{\pi' \in \Pi} P_t(\pi') \exp(-\eta\widehat{c}_t(\pi'(x_t)))\right) = \ln\left(\frac{P_t(\pi)\exp(-\eta\widehat{c}_t(\pi(x_t)))}{\tilde{P}_{t+1}(\pi)}\right)$$
$$= \ln\left(\frac{P_t(\pi)(1 - N\mu)}{P_{t+1}(\pi) - \mu}\right) - \eta\widehat{c}_t(\pi(x_t))$$
$$\geq \ln(1 - N\mu) + \ln\left(\frac{P_t(\pi)}{P_{t+1}(\pi)}\right) - \eta\widehat{c}_t(\pi(x_t))$$
$$\geq -2N\mu + \ln\left(\frac{P_t(\pi)}{P_{t+1}(\pi)}\right) - \eta\widehat{c}_t(\pi(x_t))$$

where the last step is by the fact $N\mu \leq \frac{1}{2}$ and thus $\ln(\frac{1}{1-N\mu}) = \ln(1 + \frac{N\mu}{1-N\mu}) \leq \ln(1 + 2N\mu) \leq 2N\mu$. Combining the above two displayed equations, summing over $t \in \mathcal{I}$, telescoping and rearranging gives

$$\sum_{t \in \mathcal{I}} c_t(a_t) - \widehat{c}_t(\pi(x_t)) \leq \frac{\ln(1/\mu) + 2LN\mu}{\eta} + \eta\sum_{t \in \mathcal{I}} \widehat{c}_t(a_t).$$

Taking the expectation on both sides, using the fact $\mathbb{E}_{a_t \sim p_t}[\widehat{c}_t(a_t)] \leq K$, and plugging $c_t(a) = 1 - r_t(a)$, η and μ finish the proof. ∎

Proof [Proof of Corollary 1] We first partition $[1, T]$ evenly into T/L intervals, then within each interval, further partition it into several subintervals so that $\mathcal{D}_t$ remains the same on each subinterval. Since the number of switches is at most $S - 1$, this process results in at most $T/L + S$ subintervals, each with length at most L. We can now apply Theorem 2 to each subinterval and sum up the regrets to get the claim bounds. ∎

Appendix C. Proofs for ADA-GREEDY

Before we prove the theorems, we first define some notations that facilitate the analysis. These notations are used throughout Appendix C, D, and E. In ADA-GREEDY and ADA-ILTCB, we define $\text{FLAG}_t = (t \geq T_i + L)$ or ($j > 1$ and $\text{NONSTATTEST}(t) = \textit{True}$), where i and j are the epoch and block indices t is in. This is exactly the condition that triggers the rerun of the algorithm. In all three algorithms, we let $B(i, j) \triangleq [T_i + 1, T_i + 2^{j-1} - 1]$; sometimes when i, j are already specified, we simply write B. Note that when $j = 1$, $B = [T_i + 1, T_i]$, which is an empty set. In this case, instead

of defining β_B (which is used in ADA-GREEDY and ADA-BINGREEDY) to be infinity, we let it to be zero. This just makes some analysis easier.

Below we state a few useful lemmas before proving the main theorem.

Lemma 10 *For any interval $\mathcal{I}$ such that $\Delta_{\mathcal{I}} \leq v$, we have for any sub-intervals $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathcal{I}$ and any $\pi \in \Pi$,*

$$|\mathcal{R}_{\mathcal{I}_1}(\pi) - \mathcal{R}_{\mathcal{I}_2}(\pi)| \leq v.$$

Proof The proof involves noticing for that any two rounds $s, t \in \mathcal{I}$ and $\pi \in \Pi$, $|\mathcal{R}_s(\pi) - \mathcal{R}_t(\pi)| \leq v$. This is easily seen using triangle inequality, since assuming $s < t$,

$$|\mathcal{R}_s(\pi) - \mathcal{R}_t(\pi)| \leq \sum_{\tau=s+1}^{t} |\mathcal{R}_\tau(\pi) - \mathcal{R}_{\tau-1}(\pi)| \leq \sum_{\tau \in \mathcal{I}} |\mathcal{R}_\tau(\pi) - \mathcal{R}_{\tau-1}(\pi)| \leq v.$$

The lemma is now immediate, since

$$|\mathcal{R}_{\mathcal{I}_1}(\pi) - \mathcal{R}_{\mathcal{I}_2}(\pi)| \leq \frac{1}{|\mathcal{I}_1|} \frac{1}{|\mathcal{I}_2|} \sum_{s \in \mathcal{I}_1} \sum_{t \in \mathcal{I}_2} |\mathcal{R}_s(\pi) - \mathcal{R}_t(\pi)| \leq v.$$

∎

Definition 11 (EVENT1) *Define EVENT1 to be the following event: for all $\mathcal{I} \subseteq [1, T]$ and all $\pi \in \Pi$,*

$$\left| \widehat{\mathcal{R}}_{\mathcal{I}}(\pi) - \mathcal{R}_{\mathcal{I}}(\pi) \right| \leq \beta_{\mathcal{I}}. \tag{1}$$

Recall $\widehat{r}_t(a) = \frac{r_t(a)}{p_t(a)} \mathbf{1}\{a = a_t\} \leq 1/\mu$ and $\mathbb{E}_t[\widehat{r}_t(\pi(x_t))] = \mathcal{R}_t(\pi), \mathbb{E}_t[\widehat{r}_t(\pi(x_t))^2] \leq 1/\mu$. By Freedman's inequality (Lemma 8) and a union bound, we have with probability at least $1 - \delta/2$, EVENT1 holds.

Lemma 12 *Consider an interval $\mathcal{I}$ where $|\mathcal{I}| \leq L$ and $\Delta_{\mathcal{I}} \leq v$. If EVENT1 holds, then there is at most one $t \in \mathcal{I}$ such that $\text{FLAG}_t = True$ (FLAG_t is defined at the beginning of Appendix C).*

Proof Let there be multiple such time instances. Let $t', t \in \mathcal{I}$ be two consecutive ones and $t' < t$. Note that $\text{FLAG}_t = True$ has two possible cases: $t \geq t' + L$ or ($j > 1$ and $\text{NONSTATTEST}(t) = True$). The former case cannot happen because $|\mathcal{I}| \leq L$. Now assume the latter. Let i, j be the epoch and block index at time t respectively. Define $A = [t - \ell + 1, t]$ to be the interval that makes $\text{NONSTATTEST}(t)$ return *True*, and define $B = [t' + 1, t' + 2^{j-1} - 1]$. Then $\text{NONSTATTEST}(t) = True$ implies

$$\widehat{\mathcal{R}}_A(\hat{\pi}_A) > \widehat{\mathcal{R}}_A(\hat{\pi}_B) + 2\beta_A + 2\beta_B + 4v. \tag{2}$$

By the optimality of $\hat{\pi}_B$, We have

$$\widehat{\mathcal{R}}_B(\hat{\pi}_A) \leq \widehat{\mathcal{R}}_B(\hat{\pi}_B). \tag{3}$$

Combining Eq. (2) and Eq. (3), we see that either $\pi = \hat{\pi}_A$ or $\pi = \hat{\pi}_B$ will make the following inequality hold:

$$\left|\widehat{\mathcal{R}}_A(\pi) - \widehat{\mathcal{R}}_B(\pi)\right| > \beta_A + \beta_B + 2v. \tag{4}$$

Thus,

$$\left|\mathcal{R}_A(\pi) - \mathcal{R}_B(\pi)\right| \geq \left|\widehat{\mathcal{R}}_A(\pi) - \widehat{\mathcal{R}}_B(\pi)\right| - \beta_A - \beta_B \qquad \text{(by (1))}$$
$$> \beta_A + \beta_B + 2v - \beta_A - \beta_B \geq 2v. \qquad \text{(by (4))}$$

On the other hand, by Lemma 10, we actually have $\left|\mathcal{R}_A(\pi) - \mathcal{R}_B(\pi)\right| \leq v$, which leads to a contradiction. Thus we can conclude that such t does not exist.

∎

Proof [Proof of Theorem 4] We condition on EVENT1. When this event holds true, by Lemma 12, there is at most one $t \in \mathcal{I}$ such that FLAG$_t$ is *True* (that is, rerun triggered at t). With this fact, we can focus on the case in which FLAG$_t$ is *False* for all $t \in \mathcal{I}$. If there is actually a $t' \in \mathcal{I}$ such that FLAG$_{t'}$ = *True*, we can divide $\mathcal{I}$ into $\mathcal{I}_1 \cup \{t'\} \cup \mathcal{I}_2$ and bound the regret in $\mathcal{I}_1$ and $\mathcal{I}_2$ separately. The total regret on $\mathcal{I}$ would then be bounded by their sum plus 1, which is still of the same order.

We will also use the fact (proven in Lemma 10) that by the condition $\Delta_{\mathcal{I}} \leq v$, we have for any $\mathcal{I}_1, \mathcal{I}_2 \subset \mathcal{I}$ and $\pi \in \Pi$,

$$|\mathcal{R}_{\mathcal{I}_1}(\pi) - \mathcal{R}_{\mathcal{I}_2}(\pi)| \leq v. \tag{5}$$

Let $\mathcal{I} = [s, e]$. For any $t \in \mathcal{I}$, define $\ell_t = 2^{\lfloor \log_2(t-s+1) \rfloor}$ (i.e., ℓ_t is the longest $\ell \in \{1, 2, 4, 8, \ldots\}$ such that $[t - \ell + 1, t] \subseteq \mathcal{I}$). Now focus on a specific t that is in epoch i and block j. Denote $A = [t - \ell_t + 1, t], B = [T_i + 1, T_i + 2^{j-1} - 1]$. Assuming the case described above (i.e., for all $\tau \in \mathcal{I}$, FLAG$_\tau$ = *False*), we have for $j > 1$ and any $\pi \in \Pi$,

$$\mathcal{R}_t(\pi) \leq \mathcal{R}_A(\pi) + v \leq \widehat{\mathcal{R}}_A(\pi) + \beta_A + v \qquad \text{(by (5) and (1))}$$
$$\leq \widehat{\mathcal{R}}_A(\hat{\pi}_A) + \beta_A + v \qquad \text{(by the optimality of } \hat{\pi}_A)$$
$$\leq \widehat{\mathcal{R}}_A(\hat{\pi}_B) + 3\beta_A + 2\beta_B + 5v \qquad \text{(NONSTATTEST}(t) = \textit{False})$$
$$\leq \mathcal{R}_A(\hat{\pi}_B) + 4\beta_A + 2\beta_B + 5v \leq \mathcal{R}_t(\hat{\pi}_B) + 4\beta_A + 2\beta_B + 6v. \qquad \text{(by (1) and (5))}$$

Note that $\beta_B = \mathcal{O}(\beta_A)$ because $|B| = 2^{j-1} - 1 \geq \frac{2^j - 1}{3} \geq \frac{t - T_i}{3} \geq \frac{\ell_t}{3} = \frac{|A|}{3}$.

Now using this bound for all $t \in \mathcal{I}$ and noting that there is at most one round with $j = 1$, we can bound the sum of conditional expected regrets by

$$\sum_{t \in \mathcal{I}} \mathbb{E}_t[r_t(\pi(x_t)) - r_t(a_t)] \leq \sum_{t \in \mathcal{I}} (\mathcal{R}_t(\pi_t^\star) - \mathcal{R}_t(\hat{\pi}_B) + K\mu)$$

$$\leq \mathcal{O}\left(\sum_{t \in \mathcal{I}} (v + \beta_{[t - \ell_t + 1, t]} + K\mu)\right)$$

$$= \tilde{\mathcal{O}}\left(|\mathcal{I}|v + \sum_{t \in \mathcal{I}}\left(\sqrt{\frac{\ln(N/\delta)}{\mu \ell_t}} + \frac{\ln(N/\delta)}{\mu \ell_t} + K\mu\right)\right)$$

$$= \tilde{\mathcal{O}}\left(|\mathcal{I}|v + L^{\frac{1}{6}}\sqrt{K|\mathcal{I}|\ln(N/\delta)} + L^{\frac{1}{3}}\sqrt{K\ln(N/\delta)}\right),$$

where in the last step we use the fact $|\mathcal{I}|L^{-\frac{1}{3}} \leq L^{\frac{1}{6}}\sqrt{|\mathcal{I}|}$ for $|\mathcal{I}| \leq L$. Finally, applying Hoeffding-Azuma inequality finishes the proof. ∎

Appendix D. Omitted Details for ADA-ILTCB

Optimization Problem (OP)

Given a time interval $\mathcal{I}$ and minimum probability μ, find $Q \in \Delta^\Pi$ such that for constant $B = 5 \times 10^5$:

$$\sum_{\pi \in \Pi} Q(\pi)\widehat{\text{Reg}}_{\mathcal{I}}(\pi) \leq 2BK\mu \tag{6}$$

$$\forall \pi \in \Pi : \ \widehat{V}_{\mathcal{I}}(Q, \pi) \leq 2K + \frac{\widehat{\text{Reg}}_{\mathcal{I}}(\pi)}{B\mu} \tag{7}$$

Figure 1: A subroutine for ADA-ILTCB, adapted from (Agarwal et al., 2014)

The optimization problem (OP) needed for ADA-ILTCB is included in Figure 1. It is almost identical to the one proposed in (Agarwal et al., 2014) except: 1) Instead of returning a sub-distribution, our version returns an exact distribution. However, as discussed in (Agarwal et al., 2014) this makes no real difference since given a sub-distribution which satisfies Eq. (6) and Eq. (7), one can always put all the remaining weight on the empirical best policy $\text{argmax}_\pi \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)$ to obtain a distribution that still satisfies those two constraints. 2) The constant B used in (Agarwal et al., 2014) is 100. It is also clear from the proof of (Agarwal et al., 2014) that the value of this constant does not affect the feasibility of (OP) nor the efficiency of finding the solution.

Let $d = \ln(8T^2N^2/\delta)\ln(L)$. Without loss of generality, below we assume $L \geq 4Kd$ so that $\mu = \min\{\frac{1}{2K}, \sqrt{\frac{d}{KL}}\} = \sqrt{\frac{d}{LK}}$. Indeed, if $L < 4Kd$, then the bound in Theorem 5 holds trivially since $L \leq 2\sqrt{LKd}$. The fact $d/\mu = LK\mu$ will be used frequently. We use V_t as a shorthand for $V_{\{t\}}$, that is, $V_t(Q, \pi) = \mathbb{E}_{x \sim D_t^{\mathcal{X}}}\left[\frac{1}{Q^\mu(\pi(x)|x)}\right]$.

We first state two lemmas that relates the variation of $\text{Reg}_t(\pi)$ and $V_t(Q, \pi)$ to $\bar{\Delta}$, and then two lemmas on the concentration bounds of empirical reward and empirical variance.

Lemma 13 *For any interval $\mathcal{I}$ such that $\bar{\Delta}_{\mathcal{I}} \leq v$, we have for any sub-intervals $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathcal{I}$ and any $\pi \in \Pi$,*

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

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

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

By Lemma 10, 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 and using $\Delta_{\mathcal{I}} \leq \bar{\Delta}_{\mathcal{I}}$ (Lemma 9), we get the desired bound. ∎

Lemma 14 *For any interval $\mathcal{I}$ such that $\bar{\Delta}_{\mathcal{I}} \leq v$, we have for any sub-intervals $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathcal{I}$, any distribution Q over Π, and any $\pi \in \Pi$,*

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

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

$$
\begin{aligned}
\left| V_s(Q, \pi) - V_t(Q, \pi) \right| &= \left| \mathbb{E}_{\mathcal{D}_s^{\mathcal{X}}} \left[\frac{1}{Q^\mu(\pi(x)|x)} \right] - \mathbb{E}_{\mathcal{D}_t^{\mathcal{X}}} \left[\frac{1}{Q^\mu(\pi(x)|x)} \right] \right| \\
&= \left| \int_{\mathcal{X}} (\mathcal{D}_s^{\mathcal{X}}(x) - \mathcal{D}_t^{\mathcal{X}}(x)) \frac{1}{Q^\mu(\pi(x)|x)} dx \right| \\
&\leq \frac{1}{\mu} \int_{\mathcal{X}} \left| \mathcal{D}_s^{\mathcal{X}}(x) - \mathcal{D}_t^{\mathcal{X}}(x) \right| dx \\
&\leq \frac{1}{\mu} \sum_{\tau = s+1}^{t} \| \mathcal{D}_\tau - \mathcal{D}_{\tau-1} \|_{\mathrm{TV}} \leq \frac{v}{\mu}.
\end{aligned}
$$

Therefore,

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

∎

Lemma 15 *With probability at least $1 - \delta/4$, ADA-ILTCB ensures that for all distributions $Q \in \Delta^\Pi$, all $\pi \in \Pi$, all intervals $\mathcal{I}$,*

$$\widehat{V}_{\mathcal{I}}(Q, \pi) \leq 6.4 V_{\mathcal{I}}(Q, \pi) + \frac{80LK}{|\mathcal{I}|}, \qquad V_{\mathcal{I}}(Q, \pi) \leq 6.4 \widehat{V}_{\mathcal{I}}(Q, \pi) + \frac{80LK}{|\mathcal{I}|}. \qquad (8)$$

Proof This is a consequence of the contexts being drawn independently. A similar argument of (Agarwal et al., 2014, Lemma 10) shows that with probability at least $1 - \delta/4$, the differences $\widehat{V}_{\mathcal{I}}(Q, \pi) - 6.4 V_{\mathcal{I}}(Q, \pi)$ and $V_{\mathcal{I}}(Q, \pi) - 6.4 \widehat{V}_{\mathcal{I}}(Q, \pi)$ are both bounded by

$$\frac{75 \ln(N)}{\mu^2 |\mathcal{I}|} + \frac{6.3 \ln(8T^2 N^2/\delta)}{\mu |\mathcal{I}|} \leq \frac{75LK}{|\mathcal{I}|} + \frac{6.3d}{\mu |\mathcal{I}|} = \frac{75LK}{|\mathcal{I}|} + \frac{6.3LK\mu}{|\mathcal{I}|} \leq \frac{80LK}{|\mathcal{I}|},$$

which completes the proof. ∎

Lemma 16 *With probability at least $1 - \delta/4$,* ADA-ILTCB *ensures that for all $\pi \in \Pi$ and all intervals $\mathcal{I}$,*

$$|\widehat{\mathcal{R}}_{\mathcal{I}}(\pi) - \mathcal{R}_{\mathcal{I}}(\pi)| \leq \frac{\mu}{|\mathcal{I}| \ln(L)} \sum_{t \in \mathcal{I}} V_t(Q_t, \pi) + \frac{LK\mu}{|\mathcal{I}|}. \qquad (9)$$

Proof By (Agarwal et al., 2014, Lemma 11), for any choice of $\lambda \in [0, \mu]$, we have with probability at least $1 - \delta/4$, for all $\pi \in \Pi$ and all intervals $\mathcal{I}$,

$$|\widehat{\mathcal{R}}_{\mathcal{I}}(\pi) - \mathcal{R}_{\mathcal{I}}(\pi)| \leq \frac{\lambda}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} V_t(Q_t, \pi) + \frac{\ln(8T^2 N/\delta)}{\lambda|\mathcal{I}|}.$$

Picking $\lambda = \mu/\ln(L)$ and using the fact $\ln(8T^2 N/\delta) \ln(L) \leq d$ and $d/\mu = LK\mu$ complete the proof. ∎

Definition 17 (EVENT2) *Let* EVENT2 *be the event that both Eq. (8) and (9) hold for all $\pi \in \Pi$, all intervals $\mathcal{I}$ and all $Q \in \Delta^\Pi$. This event happens with probability at least $1 - \delta/2$.*

Next we prove the following key lemma on the concentration of empirical regrets.

Lemma 18 *Conditioning on* EVENT2*, for any $\pi \in \Pi$, any interval $\mathcal{I}$ such that $|\mathcal{I}| \leq L$, $\bar{\Delta}_{\mathcal{I}} \leq v$, and there is no rerun triggered in $\mathcal{I}$ (i.e., $\forall t \in \mathcal{I}$, FLAG$_t$ = False), we have*

$$\text{Reg}_{\mathcal{I}}(\pi) \leq 2\widehat{\text{Reg}}_{\mathcal{I}}(\pi) + \frac{D_1 LK\mu}{|\mathcal{I}|} + D_2 v, \qquad \widehat{\text{Reg}}_{\mathcal{I}}(\pi) \leq 2\text{Reg}_{\mathcal{I}}(\pi) + \frac{D_1 LK\mu}{|\mathcal{I}|} + D_2 v, \quad (10)$$

where $D_1 \triangleq 2 \times 10^5$ and $D_2 \triangleq 360$.

Proof We prove the lemma by induction on the length of $\mathcal{I}$. For the base case $|\mathcal{I}| = 1$, the bounds hold trivially since both $\text{Reg}_{\mathcal{I}}(\pi)$ and $\widehat{\text{Reg}}_{\mathcal{I}}(\pi)$ are bounded by $1/\mu = LK\mu/d \leq D_1 LK\mu$. Now assuming that the statement holds for any $\mathcal{I}'$ such that $|\mathcal{I}'| \leq L' < L$, we prove below it holds for any $\mathcal{I}$ such that $|\mathcal{I}| = L' + 1$ too.

Let $\mathcal{I} = [s, e]$ belong to epoch i and block j. For every $t \in [s+1, e]$, define $\ell_t = 2^{\lfloor \log_2(t-s) \rfloor}$ (i.e., ℓ_t is the longest $\ell \in \{1, 2, 4, 8, \ldots\}$ such that $[t - \ell, t - 1] \subseteq \mathcal{I}$). Based on the induction assumption, we first prove the property $V_t(Q_t, \pi) \leq \frac{\text{Reg}_{\mathcal{I}}(\pi)}{2\mu} + \mathcal{O}\left(\frac{LK}{t-s+1} + \frac{v}{\mu}\right)$ for all π: when

$t \in [s+1, e]$,

$$V_t(Q_t, \pi) \leq V_{[t-\ell_t, t-1]}(Q_t, \pi) + \frac{v}{\mu} \qquad \text{(by Lemma 14)}$$

$$\leq 6.4 \widehat{V}_{[t-\ell_t, t-1]}(Q_t, \pi) + \frac{80LK}{\ell_t} + \frac{v}{\mu} \qquad \text{(by Eq. (8))}$$

$$\leq 263 \widehat{V}_{[T_i+1, T_i+2^{j-1}-1]}(Q_t, \pi) + \frac{7.76 \times 10^3 LK}{\ell_t} + \frac{42v}{\mu} \qquad \text{(by Line 17)}$$

$$\leq \frac{5.26 \times 10^{-4}}{\mu} \widehat{\text{Reg}}_{[T_i+1, T_i+2^{j-1}-1]}(\pi) + \frac{8.29 \times 10^3 LK}{\ell_t} + \frac{42v}{\mu} \qquad \text{(by Eq. (7) and } L \geq \ell_t)$$

$$\leq \frac{2.11 \times 10^{-3}}{\mu} \widehat{\text{Reg}}_{[t-\ell_t, t-1]}(\pi) + \frac{8.82 \times 10^3 LK}{\ell_t} + \frac{43v}{\mu} \qquad \text{(by Line 15)}$$

$$\leq \frac{4.22 \times 10^{-3}}{\mu} \text{Reg}_{[t-\ell_t, t-1]}(\pi) + \frac{9.25 \times 10^3 LK}{\ell_t} + \frac{44v}{\mu} \qquad \text{(by inductive assumption)}$$

$$\leq \frac{\text{Reg}_{\mathcal{I}}(\pi)}{2\mu} + \frac{1.9 \times 10^4 LK}{t-s+1} + \frac{45v}{\mu}; \qquad (\text{ by Lemma 13 and } t-s+1 \leq 2\ell_t)$$

$$(11)$$

when $t = s$, Eq.(11) also holds because $V_s(Q_s, \pi) \leq \frac{1}{\mu} \leq \frac{\sqrt{LK}}{t-s+1}$.

Let $\pi_{\mathcal{I}}^\star = \text{argmax}_\pi \mathcal{R}_{\mathcal{I}}(\pi)$ and $\widehat{\pi}_{\mathcal{I}} = \text{argmax}_\pi \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)$. We will now establish the inductive hypothesis. For any π, $\text{Reg}_{\mathcal{I}}(\pi) - \widehat{\text{Reg}}_{\mathcal{I}}(\pi)$ is bounded by

$$(\mathcal{R}_{\mathcal{I}}(\pi_{\mathcal{I}}^\star) - \mathcal{R}_{\mathcal{I}}(\pi)) - (\widehat{\mathcal{R}}_{\mathcal{I}}(\pi_{\mathcal{I}}^\star) - \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)) \qquad \text{(by optimality of } \widehat{\pi}_{\mathcal{I}})$$

$$\leq \frac{\mu}{|\mathcal{I}| \ln(L)} \sum_{t \in \mathcal{I}} (V_t(Q_t, \pi) + V_t(Q_t, \pi_{\mathcal{I}}^\star)) + \frac{2LK\mu}{|\mathcal{I}|} \qquad \text{(by Lemma 16)}$$

$$\leq \frac{1}{2}\text{Reg}_{\mathcal{I}}(\pi) + \left(\frac{3.8 \times 10^4 LK\mu}{|\mathcal{I}| \ln(L)} \sum_{t \in \mathcal{I}} \frac{1}{t-s+1} \right) + \frac{2LK\mu}{|\mathcal{I}|} + 90v$$

$$\text{(by Eq. (11) and } \text{Reg}_{\mathcal{I}}(\pi_{\mathcal{I}}^\star) = 0)$$

$$\leq \frac{1}{2}\text{Reg}_{\mathcal{I}}(\pi) + \frac{6 \times 10^4 LK\mu}{|\mathcal{I}|} + 90v. \qquad (12)$$

Rearranging proves the first statement of Eq. (10). Similarly, we can bound $\widehat{\text{Reg}}_{\mathcal{I}}(\pi) - \text{Reg}_{\mathcal{I}}(\pi)$ as follows:

$$(\widehat{\mathcal{R}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) - \widehat{\mathcal{R}}_{\mathcal{I}}(\pi)) - (\mathcal{R}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) - \mathcal{R}_{\mathcal{I}}(\pi)) \qquad \text{(by optimality of } \pi_{\mathcal{I}}^\star)$$

$$\leq \frac{\mu}{|\mathcal{I}| \ln(L)} \sum_{t \in \mathcal{I}} (V_t(Q_t, \pi) + V_t(Q_t, \widehat{\pi}_{\mathcal{I}})) + \frac{2LK\mu}{|\mathcal{I}|} \qquad \text{(by Lemma 16)}$$

$$\leq \frac{1}{2}(\text{Reg}_{\mathcal{I}}(\pi) + \text{Reg}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}})) + \left(\frac{3.8 \times 10^4 LK\mu}{|\mathcal{I}| \ln(L)} \sum_{t \in \mathcal{I}} \frac{1}{t-s+1} \right) + \frac{2LK\mu}{|\mathcal{I}|} + 90v \quad \text{(by Eq. (11))}$$

$$\leq \frac{1}{2}\text{Reg}_{\mathcal{I}}(\pi) + \frac{9.9 \times 10^4 LK\mu}{|\mathcal{I}|} + 180v, \qquad (13)$$

where the last step is by applying Eq. (12) to $\widehat{\pi}_{\mathcal{I}}$ and using the fact $\widehat{\mathrm{Reg}}_{\mathcal{I}}(\widehat{\pi}_{\mathcal{I}}) = 0$. Rearranging proves the second statement of Eq. (10), which completes the induction. ∎

Lemma 19 *Consider an interval $\mathcal{I}$ where $|\mathcal{I}| \leq L$ and $\bar{\Delta}_{\mathcal{I}} \leq v$. If the event* EVENT2 *holds, then there is at most one $t \in \mathcal{I}$ such that* FLAG$_t$ = *True.*

Proof If there are multiple such time instances, let $t', t \in \mathcal{I}$ be consecutive ones, and let i and j be the epoch and block indices at time t. For FLAG(t) = *True*, there are two possibilities: $t \geq t' + L$ or ($j > 1$ and NONSTATTEST(t) = *True*). The former would not happen because $|\mathcal{I}| \leq L$. Thus the latter holds. Since $j > 1$, we have $t \geq t' + 2$. By our construction, $\mathcal{I}' \triangleq [t' + 1, t - 1]$ is an interval in which no rerun is triggered, and $\bar{\Delta}_{\mathcal{I}'} \leq v$ holds. Using Lemma 18 on $\mathcal{I}'$, we have for any $1 \leq \ell \leq t - t' - 1$,

$$
\begin{aligned}
\widehat{\mathrm{Reg}}_{[t'+1,t'+2^{j-1}-1]}(\pi) &\leq 2\mathrm{Reg}_{[t'+1,t'+2^{j-1}-1]}(\pi) + \frac{2 \times 10^5 LK\mu}{2^{j-1}-1} + 360v && \text{(by Lemma 18)} \\
&\leq 2\mathrm{Reg}_{[t-\ell,t-1]}(\pi) + \frac{2 \times 10^5 LK\mu}{2^{j-1}-1} + 364v && \text{(by Lemma 13)} \\
&\leq 4\widehat{\mathrm{Reg}}_{[t-\ell,t-1]}(\pi) + \frac{8 \times 10^5 LK\mu}{\ell} + 1084v, && \\
& \qquad \text{(by Lemma 18 and } \ell \leq t - t' - 1 \leq 2^j - 2)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathrm{Reg}}_{[t-\ell,t-1]}(\pi) &\leq 2\mathrm{Reg}_{[t-\ell,t-1]}(\pi) + \frac{2 \times 10^5 LK\mu}{\ell} + 360v && \text{(by Lemma 18)} \\
&\leq 2\mathrm{Reg}_{[t'+1,t'+2^{j-1}-1]}(\pi) + \frac{2 \times 10^5 LK\mu}{\ell} + 364v && \text{(by Lemma 13)} \\
&\leq 4\widehat{\mathrm{Reg}}_{[t'+1,t'+2^{j-1}-1]}(\pi) + \frac{1 \times 10^6 LK\mu}{\ell} + 1084v, && \\
& \qquad \text{(by Lemma 18 and } \ell \leq t - t' - 1 \leq 2^j - 2)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{V}_{[t-\ell,t-1]}(Q, \pi) &\leq 6.4 V_{[t-\ell,t-1]}(Q, \pi) + \frac{80LK}{\ell} && \text{(by Eq. (8))} \\
&\leq 6.4 V_{[t'+1,t'+2^{j-1}-1]}(Q, \pi) + \frac{80LK}{\ell} + 6.4v && \text{(by Lemma 13)} \\
&\leq 41 \widehat{V}_{[t'+1,t'+2^{j-1}-1]}(Q, \pi) + \frac{1104LK}{\ell} + 6.4v. && \text{(by Eq. (8) and } \ell \leq 2^j - 2)
\end{aligned}
$$

Therefore, at time t, NONSTATTEST(t) should return *True*, which contradicts our assumption. ∎

We can now prove Theorem 5.

Proof [Proof of Theorem 5] Conditioning on EVENT2, we can focus on the case when there is no rerun in $\mathcal{I}$ (i.e., $\forall t \in \mathcal{I}$, FLAG$_t$ = *False*). This is because by Lemma 19, there is at most one $t' \in \mathcal{I}$ such that FLAG$_{t'}$ = *True*. Suppose this t' exists, we can decompose $\mathcal{I}$ into $\mathcal{I}_1 \cup \{t'\} \cup \mathcal{I}_2$, where FLAG$_t$ = *False* for all $t \in \mathcal{I}_1$ or $\mathcal{I}_2$, and then we can apply our proof to $\mathcal{I}_1$ and $\mathcal{I}_2$ separately. The regret in $\mathcal{I}$ would then be bounded by their sum plus 1, which is still of the same order.

With notation s, ℓ_t, i, j from the proof of Lemma 18, for any $t \in \mathcal{I}$ and $t \geq s + 2$, we have

$$
\sum_{\pi \in \Pi} Q_t(\pi) \mathrm{Reg}_{\mathcal{I}}(\pi)
$$

$$
\leq \sum_{\pi \in \Pi} Q_t(\pi) \mathrm{Reg}_{[t-\ell_t, t-1]}(\pi) + 2v \qquad \text{(by Lemma 13)}
$$

$$
\leq 2 \sum_{\pi \in \Pi} Q_t(\pi) \widehat{\mathrm{Reg}}_{[t-\ell_t, t-1]}(\pi) + \frac{D_1 L K \mu}{\ell_t} + (D_2 + 2)v \qquad \text{(by Lemma 18)}
$$

$$
\leq 2C_1 \sum_{\pi \in \Pi} Q_t(\pi) \widehat{\mathrm{Reg}}_{[T_i+1, T_i+2^{j-1}-1]}(\pi) + (2C_2 + D_1)\frac{L K \mu}{\ell_t} + (2C_3 + D_2 + 2)v
$$

$$
\text{(by Line 16)}
$$

$$
\leq 4B C_1 K \mu + (2C_2 + D_1)\frac{L K \mu}{\ell_t} + (2C_3 + D_2 + 2)v \qquad \text{(by Eq. (6))}
$$

$$
= \mathcal{O}\left(\frac{L K \mu}{t - s + 1} + v\right). \qquad (L \geq \ell_t \geq (t - s + 1)/2)
$$

Therefore, the sum of conditional expected regrets $\sum_{t \in \mathcal{I}} \mathbb{E}_t[r_t(\pi(x_t)) - r_t(a_t)]$ is bounded by

$$
L K \mu + (1 - K\mu) \sum_{t \in \mathcal{I}} \sum_{\pi \in \Pi} Q_t(\pi) \mathrm{Reg}_{\mathcal{I}}(\pi) = \widetilde{\mathcal{O}}(|\mathcal{I}|v + L K \mu) = \widetilde{\mathcal{O}}\left(|\mathcal{I}|v + \sqrt{L K \ln(N/\delta)}\right).
$$

The theorem now follows by an application of the Hoeffding-Azuma inequality. ∎

Appendix E. Omitted Details for ADA-BINGREEDY

In ADA-BINGREEDY, the bin length is set to $H^{\frac{1}{2}}$, and the probability of an exploration bin is $b^{-\frac{1}{2}}$. These two values are not clear before we derive the regret bound and select them optimally. In the following analysis, we will keep them as variables before reaching the final steps. Specifically, we let the bin length be H^γ (therefore Line 7 would be a for-loop from 1 to $H^{1-\gamma}$, while Line 9 from 1 to H^γ); and we let the exploration probability at Line 8 be $b^{-\theta}$.

In ADA-BINGREEDY, if an interval $\mathcal{I}$ is an subinterval of an exploration bin, by Freedman's inequality, with probability at least $1 - \delta/T^2$,

$$
\left|\widehat{\mathcal{R}}_{\mathcal{I}}(\pi) - \mathcal{R}_{\mathcal{I}}(\pi)\right| \leq \alpha_{\mathcal{I}}. \tag{14}
$$

For a general interval $\mathcal{I}$, we have with probability at least $1 - \delta/T^2$,

$$
\left|\widehat{\mathcal{R}}_{\mathcal{I}}(\pi) - \mathcal{R}_{\mathcal{I}}(\pi)\right| \leq \beta_{\mathcal{I}}. \tag{15}
$$

We now define the high probability event that is used in ADA-BINGREEDY's analysis:

Definition 20 (EVENT3) *Define* EVENT3 *to be the following event: for all $\pi \in \Pi$, all interval $\mathcal{I} \subseteq [1, T]$, Eq.(14) holds if $\mathcal{I}$ is an subinterval of an exploration bin; Eq.(15) holds if otherwise.*

A union bound over these events implies that EVENT3 holds with probability at least $1 - \delta/2$.

In this subsection, we use S' to denote the total number of epochs in the whole time horizon, and use $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_{S'}$ to denote individual epochs. Besides, we denote $K' = K \ln(N/\delta)$. We also use notations that are defined at the beginning of Appendix C.

We analyze ADA-BINGREEDY under the switching and drifting distribution settings in the following two subsections respectively.

E.1. Switching Regret

Lemma 21 *With probability at least $1 - \delta/2$, $S' \leq S$.*

Proof It suffices to prove that under EVENT3, at any time t if there is no distribution change from the start of the epoch, the algorithm will not rerun.

Let t be in epoch i and block j, and is in an exploration bin. Suppose $\mathcal{D}_{T_i+1} = \mathcal{D}_{T_i+2} = \cdots = \mathcal{D}_t$. For any ℓ such that $[t - \ell + 1, t]$ is a subset of the bin, denoting $A = [t - \ell + 1, t]$ and $B = [T_i + 1, T_i + 2^{j-1} - 1]$. When $j > 1$, we have for any π,

$$
\begin{aligned}
\widehat{\mathcal{R}}_A(\pi) &\leq \mathcal{R}_A(\pi) + \alpha_A && \text{(by (14))} \\
&= \mathcal{R}_B(\pi) + \alpha_A && \text{(the distribution does not change from } T_i \text{ to } t\text{)} \\
&\leq \widehat{\mathcal{R}}_B(\pi) + \alpha_A + \beta_B && \text{(by (15))} \\
&\leq \widehat{\mathcal{R}}_B(\hat{\pi}_B) + \alpha_A + \beta_B && \text{(by the optimality of } \hat{\pi}_B\text{)} \\
&\leq \mathcal{R}_B(\hat{\pi}_B) + \alpha_A + 2\beta_B && \text{(by (15))} \\
&= \mathcal{R}_A(\hat{\pi}_B) + \alpha_A + 2\beta_B && \text{(the distribution does not change from } T_i \text{ to } t\text{)} \\
&\leq \widehat{\mathcal{R}}_A(\hat{\pi}_B) + 2\alpha_A + 2\beta_B. && \text{(by (14))}
\end{aligned}
$$

Therefore, NONSTATTEST(t) would return *False*. Hence, conditioned on EVENT3, the algorithm ends an epoch only when there is some distribution change. This proves the lemma. ∎

Definition 22 (flat bin) *A bin $\mathcal{I}$ in epoch i and block j is called a flat bin if for all $\pi \in \Pi$ and for all $[s, e]$ such that 1) $[s, e] \subseteq \mathcal{I}$ and 2) $e - s + 1 = 2^q$ for some nonnegative integer q, the following holds (with $B(i, j) = [T_i + 1, T_i + 2^{j-1} - 1]$):*

$$
\mathcal{R}_{[s,e]}(\pi) \leq \mathcal{R}_{[s,e]}(\hat{\pi}_{(i,j)}) + 2\beta_{B(i,j)}(\pi) + 4\alpha_{[s,e]}. \tag{16}
$$

The above definition basically says that in a flat bin, $\hat{\pi}_{(i,j)}$ performs well in the sense that for any π, $\mathcal{R}_{[s,e]}(\pi) - \mathcal{R}_{[s,e]}(\hat{\pi}_{(i,j)})$ is small in all sub-intervals $[s, e]$ such that $e - s + 1 = 2^q$. Since in an exploitation bin, the learner mostly plays $\hat{\pi}_{(i,j)}$, we have the following lemma saying that the regret contributed from *flat exploitation bins* is small.

Lemma 23 ADA-BINGREEDY *always ensures the following:*

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)]\mathbf{1}\{t \text{ is in flat exploitation bins}\} \\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{K'}\left(S'^{\frac{1}{4}}T^{\frac{3}{4}} + \sqrt{ST}\right) + K'\left(\sqrt{S'T} + S\right)\right).
\end{aligned}
$$

Proof The proof will go through several stages: we sequentially calculate the regret in a bin, a block, an epoch, and then the whole time horizon; the regret in a later level is simply a summation over its previous level. Most proofs in this section are all in this form.

- **Regret in a bin.** Let $\mathcal{I}$ be a flat exploitation bin that lies in epoch i and block j. Partition $\mathcal{I}$ into $[s_1, e_1], \ldots, [s_{S_{\mathcal{I}}}, e_{S_{\mathcal{I}}}]$ such that for every $k \in [S_{\mathcal{I}}]$, $[s_k, e_k]$ is an i.i.d. interval. For every $t \in [s_k, e_k]$, define $\ell_t = 2^{\log_2 \lfloor t - s_k + 1 \rfloor}$ (that is, the longest $\ell \in \{1, 2, 4, 8, \ldots\}$ such that $[t - \ell + 1, t] \subseteq [s_k, e_k]$). By the definition of flat bin, we have for all π,

$$
\begin{aligned}
\mathcal{R}_t(\pi) &= \mathcal{R}_{[t-\ell_t+1,t]}(\pi) & ([s_k, e_k] \text{ is i.i.d.}) \\
&\leq \mathcal{R}_{[t-\ell_t+1,t]}(\hat{\pi}_{(i,j)}) + 2\beta_{B(i,j)} + 4\alpha_{[t-\ell_t+1,t]} & (\mathcal{I} \text{ is a flat bin}) \\
&= \mathcal{R}_t(\hat{\pi}_{(i,j)}) + 2\beta_{B(i,j)} + 4\alpha_{[t-\ell_t+1,t]}. & ([s_k, e_k] \text{ is i.i.d.})
\end{aligned}
$$

Therefore,

$$
\mathcal{R}_t(\pi) - \mathcal{R}_t(\hat{\pi}_{(i,j)}) \leq \widetilde{\mathcal{O}}\left(\beta_{B(i,j)} + \sqrt{\frac{K'}{\ell_t}} + \frac{K'}{\ell_t} \right) = \widetilde{\mathcal{O}}\left(\beta_{B(i,j)} + \sqrt{\frac{K'}{t - s_k + 1}} + \frac{K'}{t - s_k + 1} \right).
$$

Thus,

$$
\begin{aligned}
\sum_{t \in \mathcal{I}} &\mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)] \\
&\leq \sum_{t \in \mathcal{I}} (\mathcal{R}_t(\pi_t^\star(x_t)) - \mathcal{R}_t(\hat{\pi}_{(i,j)})) + \sum_{t \in \mathcal{I}} K\mu_t \\
&= \sum_{k=1}^{S_{\mathcal{I}}} \sum_{t=s_k}^{e_k} (\mathcal{R}_t(\pi_t^\star(x_t)) - \mathcal{R}_t(\hat{\pi}_{(i,j)})) + \sum_{t \in \mathcal{I}} K\mu_t \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S_{\mathcal{I}}} \sqrt{K'(e_k - s_k + 1)} + K' + \sum_{t \in \mathcal{I}} (\beta_{B(i,j)} + K\mu_t) \right) \\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{K' S_{\mathcal{I}} |\mathcal{I}|} + K' S_{\mathcal{I}} + \sum_{t \in \mathcal{I}} (\beta_{B(i,j)} + K\mu_t) \right),
\end{aligned}
$$

where in the last inequality, we use Cauchy-Schwarz inequality.

- **Regret in a block.** Now we compute the regret contributed from flat explotation bins in a block $\mathcal{J}$ whose epoch and block indices are i and j respectively. Assume there are Γ bins in $\mathcal{J}$. Then $|\mathcal{J}| \leq 2^{j-1}$ and $\Gamma \leq 2^{(j-1)(1-\gamma)}$ by the algorithm. Let $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_\Gamma$ be the bins in $\mathcal{J}$, we have $\sum_{b=1}^{\Gamma} S_{\mathcal{I}_b} \leq S_{\mathcal{J}} + \Gamma$ (because the boundaries between bins can cut a stationary interval into two).

By our conclusion at the previous stage,

$$\sum_{t \in \mathcal{J}} \mathbb{E}_t[(r_t(\pi_t^\star(x_t)) - r_t(a_t))]\mathbf{1}\{t \text{ is in flat exploitation bins}\}$$

$$\leq \sum_{b=1}^{\Gamma} \widetilde{\mathcal{O}}\left(\sqrt{K' S_{\mathcal{I}_b}|\mathcal{I}_b|} + K' S_{\mathcal{I}_b}\right) + \sum_{t \in \mathcal{J}} \widetilde{\mathcal{O}}(\beta_{B(i,j)} + K\mu_t)$$

$$\leq \widetilde{\mathcal{O}}\left(\sqrt{K'(S_{\mathcal{J}} + \Gamma)|\mathcal{J}|} + K'(S_{\mathcal{J}} + \Gamma)\right) + \sum_{t \in \mathcal{J}} \widetilde{\mathcal{O}}(\beta_{B(i,j)} + K\mu_t) \qquad \text{(Cauchy-Schwarz)}$$

$$\leq \widetilde{\mathcal{O}}\left(\sqrt{K'}\left(\sqrt{S_{\mathcal{J}}|\mathcal{J}|} + 2^{(j-1)(1-\frac{\gamma}{2})}\right) + K'\left(S_{\mathcal{J}} + 2^{(j-1)(1-\gamma)}\right)\right) + \sum_{t \in \mathcal{J}} \widetilde{\mathcal{O}}(\beta_{B(i,j)} + K\mu_t).$$
$$(\Gamma \leq 2^{(j-1)(1-\gamma)} \text{ and } |\mathcal{J}| \leq 2^{j-1})$$

- **Regret in an epoch.** Now we compute the regret in an epoch $\mathcal{E}$. There are $\lceil \log_2(1 + |\mathcal{E}|) \rceil$ blocks in the epoch $\mathcal{E}$, and we denote them by $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_{\lceil \log_2(1+|\mathcal{E}|) \rceil}$. Similarly, we have $\sum_{j=1}^{\lceil \log_2(1+|\mathcal{E}|) \rceil} S_{\mathcal{J}_j} \leq S_{\mathcal{E}} + \lceil \log_2(1 + |\mathcal{E}|) \rceil$. Summing up the regret in individual blocks and again using Cauchy-Schwarz inequality, we get

$$\sum_{t \in \mathcal{E}} \mathbb{E}_t[(r_t(\pi_t^\star(x_t)) - r_t(a_t))]\mathbf{1}\{t \text{ is in flat exploitation bins}\}$$

$$\leq \widetilde{\mathcal{O}}\left(\sqrt{K'}\left(\sqrt{S_{\mathcal{E}}|\mathcal{E}|} + \sum_{j=1}^{\lceil \log_2(1+|\mathcal{E}|) \rceil} 2^{(j-1)(1-\frac{\gamma}{2})}\right) + K'\left(S_{\mathcal{E}} + \sum_{j=1}^{\lceil \log_2(1+|\mathcal{E}|) \rceil} 2^{(j-1)(1-\gamma)}\right)\right)$$

$$+ \widetilde{\mathcal{O}}\left(\sum_{t \in \mathcal{E}}(\sqrt{K'}(t - T_i)^{-\frac{1}{3}} + K'(t - T_i)^{-\frac{2}{3}})\right)$$

$$= \widetilde{\mathcal{O}}\left(\sqrt{K'}\left(\sqrt{S_{\mathcal{E}}|\mathcal{E}|} + |\mathcal{E}|^{1-\frac{\gamma}{2}} + |\mathcal{E}|^{\frac{2}{3}}\right) + K'\left(S_{\mathcal{E}} + |\mathcal{E}|^{1-\gamma} + |\mathcal{E}|^{\frac{1}{3}}\right)\right)$$

$$= \widetilde{\mathcal{O}}\left(\sqrt{K'}\left(\sqrt{S_{\mathcal{E}}|\mathcal{E}|} + |\mathcal{E}|^{\frac{3}{4}}\right) + K'\left(S_{\mathcal{E}} + |\mathcal{E}|^{\frac{1}{2}}\right)\right)$$

- **Regret in the whole time horizon.** Finally, we sum the bound over epochs and use Hölder's inequality. Again, we have $\sum_{i=1}^{S'} S_{\mathcal{E}} \leq S + S'$.

$$\sum_{t=1}^{T} \mathbb{E}_t[(r_t(\pi_t^\star(x_t)) - r_t(a_t))]\mathbf{1}\{t \text{ is in flat exploitation bins}\}$$

$$\leq \widetilde{\mathcal{O}}\left(\sqrt{K'}\left(\sqrt{(S + S')T} + S'^{\frac{1}{4}}T^{\frac{3}{4}}\right) + K'(S + S' + S'^{\frac{1}{2}}T^{\frac{1}{2}})\right)$$

$$\leq \widetilde{\mathcal{O}}\left(\sqrt{K'}\left(S'^{\frac{1}{4}}T^{\frac{3}{4}} + \sqrt{ST}\right) + K'\left(\sqrt{S'T} + S\right)\right).$$

$\blacksquare$

Given we have low regret in flat exploitation bins, in the following two lemmas we bound the number of rounds in non-flat bins or exploration bins.

Lemma 24 *With probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \mathbf{1}\{t \text{ is in exploration bins}\} \leq \widetilde{\mathcal{O}}\left(S'^{\frac{1}{4}} T^{\frac{3}{4}}\right).$$

Proof

- **Regret in a block.** We first look at a block $\mathcal{J}$ whose block index is j. Recall that in this block, bin length is set to $2^{(j-1)\gamma}$. Conditioned on all history before $\mathcal{J}$, in $\mathcal{J}$ we have

$$\sum_{b=1}^{2^{(j-1)(1-\gamma)}} \mathbb{E}_{\text{bin}(b)}[\mathbf{1}\{\text{bin } b \text{ is exploration}\}] \leq \sum_{b=1}^{2^{(j-1)(1-\gamma)}} b^{-\theta} \leq \mathcal{O}(2^{(j-1)(1-\gamma)(1-\theta)}),$$

where we slightly overload the notation, using $\mathbb{E}_{\text{bin}(b)}$ to denote that expectation conditioned on all history before bin b. Applying Hoeffding-Azuma's inequality, with probability at lest $1 - \delta/(T \log_2 T)$, the number of exploration bins in $\mathcal{J}$ is upper bound by $\widetilde{\mathcal{O}}(2^{(j-1)(1-\gamma)(1-\theta)} + 2^{\frac{1}{2}(j-1)(1-\gamma)})$. In other words,

$$\sum_{t \in \mathcal{J}} \mathbf{1}\{t \text{ is in exploration bins}\} \leq 2^{(j-1)\gamma} \times \widetilde{\mathcal{O}}(2^{(j-1)(1-\gamma)(1-\theta)} + 2^{\frac{1}{2}(j-1)(1-\gamma)})$$

$$= \widetilde{\mathcal{O}}(2^{(j-1)(1-\theta+\gamma\theta)} + 2^{\frac{1}{2}(j-1)(1+\gamma)}).$$

Using a union bound, we know that with probability $1 - \delta$, the above bound holds for all j and all blocks with index j in the whole time horizon. The $T \log_2 T$ factor is because there can be at most $\log_2 T$ different j's and at most T blocks with index j. We call this EVENT4. In the following stages, we condition on EVENT4.

- **Regret in an epoch.** Now sum the bound in the previous stage over blocks in an epoch $\mathcal{E}$. Conditioning on EVENT4, we have

$$\sum_{t \in \mathcal{E}} \mathbf{1}\{t \text{ is in exploration bins}\} \leq \widetilde{\mathcal{O}}\left(\sum_{j=1}^{\lceil \log_2(1+|\mathcal{E}|) \rceil} 2^{(j-1)(1-\theta+\gamma\theta)} + 2^{\frac{1}{2}(j-1)(1+\gamma)} \right)$$

$$= \widetilde{\mathcal{O}}\left(|\mathcal{E}|^{1-\theta+\gamma\theta} + |\mathcal{E}|^{\frac{1}{2}+\frac{1}{2}\gamma} \right) = \widetilde{\mathcal{O}}\left(|\mathcal{E}|^{\frac{3}{4}} \right).$$

- **Regret in the whole time horizon.** Finally, we sum over epochs in the whole time horizon and use union bound. Conditioning on EVENT4, we have

$$\sum_{t=1}^{T} \mathbf{1}\{t \text{ is in exploration bins}\} \leq \widetilde{\mathcal{O}}\left(\sum_{i=1}^{S'} |\mathcal{E}_i|^{\frac{3}{4}} \right) \leq \widetilde{\mathcal{O}}\left(S'^{\frac{1}{4}} T^{\frac{3}{4}} \right),$$

where in the final step we use Hölder's inequality.

$\blacksquare$

Lemma 25 *With probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \mathbf{1}\{t \text{ is in non-flat bins}\} \leq \tilde{\mathcal{O}}\left(S'^{\frac{1}{4}} T^{\frac{3}{4}}\right).$$

Proof

- **Regret in an epoch.** Let $\mathcal{I}$ be a non-flat bin whose epoch and block indices are i and j. Then there exists some $[s, e] \subseteq \mathcal{I}$ such that $e - s + 1 = 2^q$ and

$$\mathcal{R}_{[s,e]}(\pi) > \mathcal{R}_{[s,e]}(\hat{\pi}_{(i,j)}) + 2\beta_{B(i,j)}(\pi) + 4\alpha_{[s,e]}.$$

Note that this can only holds for $j > 1$. Furthermore, if EVENT3 holds and $\mathcal{I}$ happens to be an exploration bin, then we have

$$\left|\mathcal{R}_{[s,e]}(\pi) - \widehat{\mathcal{R}}_{[s,e]}(\pi)\right| \leq \alpha_{[s,e]},$$
$$\left|\mathcal{R}_{[s,e]}(\hat{\pi}_{(i,j)}) - \widehat{\mathcal{R}}_{[s,e]}(\hat{\pi}_{(i,j)})\right| \leq \alpha_{[s,e]}.$$

Combining the three inequalities, we get

$$\widehat{\mathcal{R}}_{[s,e]}(\pi) > \widehat{\mathcal{R}}_{[s,e]}(\hat{\pi}_{(i,j)}) + 2\beta_{B(i,j)}(\pi) + 2\alpha_{[s,e]}.$$

This event will make NONSTATTEST$(e) = $ *True*, which then triggers the rerun. The above argument indicates that as long as EVENT3 holds, the non-flat bins in an epoch would only include the first non-flat exploration bins (in which the whole epoch ends) and all non-flat exploitation bins that appear before it. Therefore, the key is to bound the number of non-flat exploitation bins that occur before the first non-flat exploration bin.

For an epoch $\mathcal{E}$, let j^* denote the last block index in it. Define X to be the number of non-flat exploitation bins in $\mathcal{E}$ that appear before the first non-flat exploration bin. Note the following two facts: 1) the decision for a bin to be exploration or exploitation is independent of its flatness, 2) a bin with index b is exploitation with probability $1 - b^{-\theta} \leq 1 - p_{\min}$, where $p_{\min} \triangleq 2^{-(j^*-1)(1-\gamma)\theta}$. Therefore, the probability $\Pr\{X > x\}$ is upper bounded by $(1 - p_{\min})^x$. This is because when $X > x$, the first x non-flat bins in the epoch all need to be exploitation bins. Picking x to be $\frac{\ln(2T/\delta)}{p_{\min}} = 2^{(j^*-1)(1-\gamma)\theta} \ln(2T/\delta) \leq |\mathcal{E}|^{(1-\gamma)\theta} \ln(2T/\delta)$, we get $\Pr\{X > x\} \leq (1 - p_{\min})^x \leq (1/e)^{\ln(2T/\delta)} = \frac{\delta}{2T}$.

Define EVENT5 to be that in every epoch, the quantity X is smaller than $|\mathcal{E}|^{(1-\gamma)\theta} \ln(2T/\delta)$. Since there are at most T epochs, a union bound guarantees that EVENT5 holds with probability at least $1 - \delta/2$.

Thus, when EVENT3 and EVENT5 both hold, we have

$$\sum_{t \in \mathcal{E}} \mathbf{1}\{t \text{ is in non-flat bins}\} \leq |\mathcal{E}|^\gamma \times \tilde{\mathcal{O}}\left(|\mathcal{E}|^{(1-\gamma)\theta}\right) = \tilde{\mathcal{O}}(|\mathcal{E}|^{\frac{3}{4}})$$

because the bin length is at most $|\mathcal{E}|^\gamma$.

- **Regret in the whole time horizon.** Finally we sum this over epochs and use union bound. From the above discussions, with probability at least $1-(1-\Pr(\text{EVENT3}))-(1-\Pr(\text{EVENT5})) \geq 1-\delta$,

$$\sum_{t=1}^{T} \mathbf{1}\{t \text{ is in non-flat bins}\} \leq \widetilde{\mathcal{O}}\left(\sum_{i=1}^{S'}|\mathcal{E}_i|^{\frac{3}{4}}\right) \leq \widetilde{\mathcal{O}}(S'^{\frac{1}{4}}T^{\frac{3}{4}}).$$

$\blacksquare$

Proof [Proof of Theorem 7 (Part I: switching regret)] Combining Lemma 21, 23, 24, and 25, we see that with probability at least $1 - 5\delta/2$,

$$\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)] \leq \sum_{t=1}^{T} \Big(\mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)]\mathbf{1}\{t \text{ is in flat exploitation bins}\}$$

$$+ \mathbf{1}\{t \text{ is in exploration bins}\} + \mathbf{1}\{t \text{ is in non-flat bins}\}\Big)$$

$$\leq \widetilde{\mathcal{O}}\left(\sqrt{K'}S^{\frac{1}{4}}T^{\frac{3}{4}} + K'\sqrt{ST}\right).$$

Applying Hoeffding-Azuma inequality shows that with probability at least $1 - 3\delta$,

$$\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t) = \widetilde{\mathcal{O}}\left(\sqrt{K'}S^{\frac{1}{4}}T^{\frac{3}{4}} + K'\sqrt{ST}\right).$$

$\blacksquare$

E.2. Dynamic Regret

Lemma 26 *With probability $1 - \delta/2$, $S' \leq \widetilde{\mathcal{O}}(1 + K'^{-\frac{2}{5}}\Delta^{\frac{4}{5}}T^{\frac{1}{5}})$.*

Proof Suppose that EVENT3 holds. When the NONSTATTEST(t) returns *True* at some t in epoch i and block j, we have (let $A = [t - \ell + 1, t], B = B(i,j) = [T_i + 1, T_i + 2^{j-1} - 1]$):

$$\widehat{\mathcal{R}}_A(\hat{\pi}_A) > \widehat{\mathcal{R}}_A(\hat{\pi}_B) + 2\beta_B + 4\alpha_A.$$

By the optimality of $\hat{\pi}_B$, we have

$$\widehat{\mathcal{R}}_B(\hat{\pi}_A) \leq \widehat{\mathcal{R}}_B(\hat{\pi}_B).$$

The above two inequalities indicate for either $\pi = \hat{\pi}_A$ or $\pi = \hat{\pi}_B$,

$$\left|\widehat{\mathcal{R}}_A(\pi) - \widehat{\mathcal{R}}_B(\pi)\right| > \beta_B + 2\alpha_A.$$

Since EVENT3 holds,

$$\left|\widehat{\mathcal{R}}_A(\pi) - \mathcal{R}_A(\pi)\right| \leq \alpha_A,$$
$$\left|\widehat{\mathcal{R}}_B(\pi) - \mathcal{R}_B(\pi)\right| \leq \beta_B.$$

Combining the above three inequalities, we get

$$\left|\mathcal{R}_A(\pi) - \mathcal{R}_B(\pi)\right| > \alpha_A.$$

Since $\left|\mathcal{R}_A(\pi) - \mathcal{R}_B(\pi)\right| \leq \Delta_{\mathcal{E}}$ and $\alpha_A = \Omega\left(\sqrt{\frac{K'}{\ell}}\right) = \Omega(K'^{\frac{1}{2}}|\mathcal{E}|^{-\frac{\gamma}{2}})$, we have $\Delta_{\mathcal{E}} \geq \Omega(K'^{\frac{1}{2}}|\mathcal{E}|^{-\frac{\gamma}{2}})$. Now invoke this lower bound for all epochs in which rerun has been triggered (i.e., $\mathcal{E}_1, \ldots, \mathcal{E}_{S'-1}$). By Hölder's inequality,

$$\begin{aligned}
S' - 1 &\leq \left(\sum_{i=1}^{S'-1} |\mathcal{E}_i|^{-\frac{\gamma}{2}}\right)^{\frac{2}{2+\gamma}} \left(\sum_{i=1}^{S'-1} |\mathcal{E}_i|\right)^{\frac{\gamma}{2+\gamma}} \\
&\leq \tilde{\mathcal{O}}\left(K'^{-\frac{1}{2+\gamma}} \left(\sum_{i=1}^{S'-1} \Delta_{\mathcal{E}_i}\right)^{\frac{2}{2+\gamma}} T^{\frac{\gamma}{2+\gamma}}\right) \\
&\leq \tilde{\mathcal{O}}\left(K'^{-\frac{1}{2+\gamma}} \Delta^{\frac{2}{2+\gamma}} T^{\frac{\gamma}{2+\gamma}}\right) \\
&= \tilde{\mathcal{O}}\left(K'^{-\frac{2}{5}} \Delta^{\frac{4}{5}} T^{\frac{1}{5}}\right).
\end{aligned}$$

∎

Lemma 27 ADA-BINGREEDY *always ensures the following*

$$\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)]\mathbf{1}\{t \text{ is in flat-exploitation bins}\} \leq \widetilde{\mathcal{O}}\left(K'\Delta^{\frac{1}{3}}T^{\frac{2}{3}} + K'S'^{\frac{1}{4}}T^{\frac{3}{4}}\right).$$

Proof

- **Regret in a bin.** If $\mathcal{I}$ is an flat exploitation bin in epoch i and block j, then for all $[s, e] \subseteq \mathcal{I}$ such that $e - s + 1 = 2^q$, we have for all π,

$$\mathcal{R}_{[s,e]}(\pi) \leq \mathcal{R}_{[s,e]}(\hat{\pi}_{(i,j)}) + 2\beta_{B(i,j)} + 4\alpha_{[s,e]},$$

which implies (by expanding the definition of $\mathcal{R}_{[s,e]}(\pi)$),

$$\begin{aligned}
\sum_{t=s}^{e} \mathbb{E}_t[r_t(\pi(x_t)) - r_t(a_t)] &\leq \sum_{t=s}^{e} (\mathcal{R}_t(\pi) - \mathcal{R}_t(\hat{\pi}_{(i,j)}) + K\mu_t) \\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{K'(e - s + 1)} + K' + \sum_{t=s}^{e}(\beta_{B(i,j)} + K\mu_t)\right). \quad (17)
\end{aligned}$$

Now we divide the whole bin into intervals of length $L' = 2^q$ for some integer q. Then we can use Lemma 6 to relate the dynamic regret in the whole bin to the sum of interval regret against a fixed policy on each of the intervals (that is, Eq. (17)).

One subtle issue is that there might be one interval (the last one) whose length is less than L'. This interval can be further divided into no more than $\log_2 L'$ subintervals whose length are all

of 2's powers. As a whole, there are no more than $\frac{|\mathcal{I}|}{L'} + \log_2 L'$ intervals each of length no more than L'. By Lemma 6 and Eq. (17), we have

$$\sum_{t \in \mathcal{I}} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)] \le \widetilde{\mathcal{O}}\left(\left(\frac{|\mathcal{I}|}{L'} + \log_2 L' \right) \left(\sqrt{K'L'} + K' \right) + L'\Delta_\mathcal{I} + \sum_{t \in \mathcal{I}} (\beta_{B(i,j)} + K\mu_t) \right).$$

Picking $L' = \min\left\{ 2^{\lfloor \log_2 |\mathcal{I}| \rfloor}, 2^{\left\lfloor \frac{2}{3} \log_2(|\mathcal{I}|/\Delta_\mathcal{I}) \right\rfloor} \right\}$, the right-hand side is further bounded by

$$\widetilde{\mathcal{O}}\left(K'|\mathcal{I}|^{\frac{2}{3}}\Delta_\mathcal{I}^{\frac{1}{3}} + K'|\mathcal{I}|^{\frac{1}{2}} + \sum_{t \in \mathcal{I}} (\beta_{B(i,j)} + K\mu_t) \right)$$

- **Regret in an epoch.** Next, we sum the regret over flat exploitation bins in an epoch. Note there are at most $\widetilde{\mathcal{O}}(|\mathcal{E}|^{1-\gamma})$ bins in an epoch $\mathcal{E}$. Using Hölder's inequality, we have

$$\sum_{t \in \mathcal{E}} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)]\mathbf{1}\{t \text{ is in flat exploitation bins}\}$$

$$\le \widetilde{\mathcal{O}}\left(K'|\mathcal{E}|^{\frac{2}{3}}\Delta_\mathcal{E}^{\frac{1}{3}} + K'|\mathcal{E}|^{1-\frac{\gamma}{2}} + K'^{\frac{1}{2}}|\mathcal{E}|^{\frac{2}{3}} \right)$$

$$= \widetilde{\mathcal{O}}\left(K'|\mathcal{E}|^{\frac{2}{3}}\Delta_\mathcal{E}^{\frac{1}{3}} + K'|\mathcal{E}|^{\frac{3}{4}} \right).$$

- **Regret in the whole time horizon.** Summing over epochs and using Hölder's inequality, we get

$$\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)]\mathbf{1}\{t \text{ is in flat exploitation bins}\} = \widetilde{\mathcal{O}}\left(K'\Delta^{\frac{1}{3}}T^{\frac{2}{3}} + K'S'^{\frac{1}{4}}T^{\frac{3}{4}} \right).$$

$\blacksquare$

Proof [Proof of Theorem 7 (Part II: dynamic regret)] Combining Lemma 26, 27, 24, and 25, we see that with probability at least $1 - 5\delta/2$,

$$\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)] \le \widetilde{\mathcal{O}}\left(K'\Delta^{\frac{1}{3}}T^{\frac{2}{3}} + K'\left(1 + \Delta^{\frac{4}{5}}T^{\frac{1}{5}} \right)^{\frac{1}{4}}T^{\frac{3}{4}} \right) \le \widetilde{\mathcal{O}}\left(K'\Delta^{\frac{1}{5}}T^{\frac{4}{5}} + K'T^{\frac{3}{4}} \right).$$

Applying Hoeffding-Azuma inequality shows that with probability at least $1 - 3\delta$,

$$\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t) = \widetilde{\mathcal{O}}\left(K'\Delta^{\frac{1}{5}}T^{\frac{4}{5}} + K'T^{\frac{3}{4}} \right).$$

The theorem finally follows by a union bound combining the switching regret bound and the dynamic regret bound we have proven. $\blacksquare$

Appendix F. Omitted Proofs in Section 4

Proof [of Lemma 6] It suffices to show that for any $i \in [n]$,

$$\sum_{t \in \mathcal{I}_i} \mathbb{E}_t \left[r_t(\pi_t^\star(x_t)) - r_t(a_t) \right] \leq \sum_{t \in \mathcal{I}_i} \mathbb{E}_t \left[r_t(\pi_{s_i}^\star(x_t)) - r_t(a_t) \right] + 2|\mathcal{I}_i| \Delta_{\mathcal{I}_i}$$

The theorem follows by summing up the regrets over all intervals.

Indeed, one can rewrite the regret as follows:

$$\sum_{t \in \mathcal{I}_i} \mathbb{E}_t \left[r_t(\pi_t^\star(x_t)) - r_t(a_t) \right] = \sum_{t \in \mathcal{I}_i} \mathbb{E}_t \left[r_t(\pi_{s_i}^\star(x_t)) - r_t(a_t) \right] + \sum_{t \in \mathcal{I}_i} \mathbb{E}_t \left[r_t(\pi_t^\star(x_t)) - r_t(\pi_{s_i}^\star(x_t)) \right]$$

$$= \sum_{t \in \mathcal{I}_i} \mathbb{E}_t \left[r_t(\pi_{s_i}^\star(x_t)) - r_t(a_t) \right] + \sum_{t \in \mathcal{I}_i} \left(\mathcal{R}_t(\pi_t^\star) - \mathcal{R}_t(\pi_{s_i}^\star) \right).$$

The last term can be further decomposed as:

$$\sum_{t \in \mathcal{I}_i} \left(\mathcal{R}_{s_i}(\pi_t^\star) - \mathcal{R}_{s_i}(\pi_{s_i}^\star) + \sum_{\tau=s_i+1}^{t} \left(\mathcal{R}_\tau(\pi_t^\star) - \mathcal{R}_{\tau-1}(\pi_t^\star) \right) + \sum_{\tau=s_i+1}^{t} \left(\mathcal{R}_{\tau-1}(\pi_{s_i}^\star) - \mathcal{R}_\tau(\pi_{s_i}^\star) \right) \right)$$

where $\mathcal{R}_{s_i}(\pi_t^\star) \leq \mathcal{R}_{s_i}(\pi_{s_i}^\star)$ by definition and the rest is bounded by $2\Delta_{\mathcal{I}_i}$. This finishes the proof. ∎

Proof [of Corollary 3] The proof of Theorem 4 shows that with probability at least $1 - \delta/2$, ADA-GREEDY ensures that for any interval $\mathcal{I}$ such that $|\mathcal{I}| \leq L$ and $\Delta_{\mathcal{I}} \leq L^{-1/3}$, we have $\sum_{t \in \mathcal{I}} \mathbb{E}_t[r_t(\pi(x_t)) - r_t(a_t)] \leq \widetilde{\mathcal{O}} \left(|\mathcal{I}|L^{-\frac{1}{3}} + L^{\frac{1}{6}} \sqrt{|\mathcal{I}|K \ln(N/\delta)} + K \ln(N/\delta) \right)$ for any π. We can thus first partition $[1, T]$ evenly into T/L intervals, then within each interval, further partition it sequentially into several largest subintervals so that for each of them the variation is at most v. Since the total variation is Δ, it is clear that this results in at most $S' \leq T/L + \Delta/L^{-1/3}$ subintervals (denote them as $\mathcal{I}_1, \ldots, \mathcal{I}_{S'}$), each of which satisfies the conditions of Theorem 4. Using Lemma 6, we get

$$\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)] \leq \sum_{i=1}^{S'} \widetilde{\mathcal{O}} \left(|\mathcal{I}_i|L^{-\frac{1}{3}} + L^{\frac{1}{6}} \sqrt{|\mathcal{I}_i|K \ln(N/\delta)} + K \ln(N/\delta) + |\mathcal{I}_i|\Delta_{\mathcal{I}_i} \right)$$

$$\leq \widetilde{\mathcal{O}} \left(\left(\frac{T}{L^{1/3}} + L^{\frac{1}{6}} \sqrt{S'T} + S' \right) K \ln(N/\delta) \right)$$

$$\leq \widetilde{\mathcal{O}} \left(\left(\frac{T}{L^{1/3}} + L^{\frac{1}{3}} \sqrt{\Delta T} + \frac{T}{L} + \Delta L^{\frac{1}{3}} \right) K \ln(N/\delta) \right)$$

$$\leq \widetilde{\mathcal{O}} \left(\left(\frac{T}{L^{1/3}} + L^{\frac{1}{3}} \sqrt{\Delta T} \right) K \ln(N/\delta) \right),$$

where in the second inequality we use Cauchy-Schwarz inequality and in the last one we use the fact $\Delta \leq T$. Finally using Hoeffding-Azuma inequality leads to the claimed bound. ∎

Proof [of Corollary 4] The proof follows the same arguments as in Corollary 3 except that now the interval regret is bounded by $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{I}|}{\sqrt{L}} + \sqrt{LK\ln(N/\delta)}\right)$, and $S' \leq T/L + \bar{\Delta}/L^{-1/2}$. Thus,

$$\sum_{t=1}^{T} \mathbb{E}_t[r_t(\pi_t^\star(x_t)) - r_t(a_t)] \leq \widetilde{\mathcal{O}}\left(\frac{T}{\sqrt{L}} + S'\sqrt{LK\ln(N/\delta)} + \bar{\Delta}L\right)$$

$$\leq \widetilde{\mathcal{O}}\left(\left(\frac{T}{\sqrt{L}} + \bar{\Delta}L\right)K\ln(N/\delta)\right).$$

Using Hoeffding-Azuma inequality gives the bound. ∎

Appendix G. Omitted Details for Corralling BISTRO+

Algorithm 5 Corralling BISTRO+
---

1 **Input**: Contexts $x_1, \ldots, x_T$ and parameter L

2 Define $\gamma = 1/T, \beta = e^{\frac{1}{\ln T}}, \eta = \min\{\frac{1}{810}, \sqrt{T/\ln N}/(LK)\}, M = \lceil T\eta \rceil$

3 Initialize $m = 1, \eta_1(i) = \eta, \rho_1(i) = 2M$ for all $i \in [M], w_1 = \bar{w}_1 = \frac{1}{M}, q_1 \in \Delta^M$ s.t. $q_1(1) = 1$

4 Initialize $\mathcal{B}_1$, a new copy of BISTRO+

5 **for** $t = 1$ **to** T **do**

6   Receive suggested action a_t^i from base algorithm $\mathcal{B}_i$ for each $i \in [m]$

7   Sample $i_t \sim q_t$, play $a_t = a_t^{i_t}$, receive reward $r_t(a_t)$

8   Construct estimated losses $\ell_t(i) = \frac{1 - r_t(a_t)}{q_t(i_t)}\mathbf{1}\{i = i_t\} + (1 - r_t(a_t))\mathbf{1}\{i > m\}, \forall i \in [M]$

9   Send feedback $\ell_t(i)$ to $\mathcal{B}_i$ for each $i \in [m]$

10   Compute $w_{t+1} \in \Delta^M$ s.t.

$$\frac{1}{w_{t+1}(i)} = \frac{1}{w_t(i)} + \eta_t(i)(\ell_t(i) + z_t(i) - \lambda)$$

     where λ is a normalization factor and $z_t(i) = 6\eta_t(i)w_t(i)(\ell_t(i) - (1 - r_t(a_t)))^2$

11   Set $\bar{w}_{t+1} = (1 - \gamma)w_{t+1} + \gamma\frac{1}{M}$

12   **for** $i = 1$ **to** M **do**

13   **if** $\frac{1}{\bar{w}_{t+1}(i)} > \rho_t(i)$ **then** set $\rho_{t+1}(i) = \frac{2}{\bar{w}_{t+1}(i)}, \eta_{t+1}(i) = \beta\eta_t(i)$

14   **else** set $\rho_{t+1}(i) = \rho_t(i), \eta_{t+1}(i) = \eta_t(i)$

15   **if** t is a multiple of $\lceil T/M \rceil$ **then**

16   Update $m \leftarrow m + 1$

17   Initialize $\mathcal{B}_m$, a new copy of BISTRO+

18   Set $q_{t+1}(i) = \frac{\bar{w}_{t+1}(i)}{\sum_{j=1}^{m} \bar{w}_{t+1}(j)}, \forall i \in [m]$

We describe the idea of using CORRAL with BISTRO+ as base algorithms (see Algorithm 5 for the pseudocode). Conceptually we always maintain M base algorithms, and use CORRAL almost in a black-box manner as in (Agarwal et al., 2017). However, crucially the i-th copy of the base algorithm only starts after the end of round $(i - 1)\lceil T/M \rceil$, in order to provide regret guarantee

starting from that round (or close to that round). Therefore, the extra work here is to make sure CORRAL does not pick algorithms that have not started, and also to come up with "virtual rewards" for algorithms before they start.

More concretely, at each time we maintain $m \leq M$ copies of the base algorithm and a distribution q_t over them (note that although q_t is in the simplex $\Delta^M := \{q \in \mathbb{R}_+^M : \sum_{i=1}^M q(i) = 1\}$, the algorithm always ensure $q_t(i) = 0$, $\forall i > m$). First we receive suggested actions a_t^i from each base algorithm $\mathcal{B}_i$. Then we sample a base algorithm $i_t \sim q_t$ and play according to its action, that is, $a_t = a_t^{i_t}$. After receiving its reward $r_t(a_t)$ (or equivalently its cost $1 - r_t(a_t)$), we construct estimated loss for each of the M algorithms: for algorithms that have started, this is simply the importance weighted loss; for algorithms that have not started, this is the actual loss of the picked action (see Line 8). Next, we send the estimated losses to the m algorithms that have started, and update several variables that CORRAL itself maintains, including the distributions w_t and $\bar{w}_t$ and the thresholds ρ_t (Line 10 to 14). Finally, we re-normalize the weights $\bar{w}_{t+1}$ over the started algorithms (including possibly a newly started one) to obtain q_{t+1} and proceed to the next round.

Another additional difference from the original CORRAL is the way we update w_t (Line 10). Here we follow the improved version proposed by Wei and Luo (2018) and incorporate an extra correction term z_t into the loss vector ℓ_t. In the original CORRAL z_t is simply the zero vector. However, with this more carefully chosen z_t we can eventually improve the bound, replacing some dependence on T by L, as shown in our proof.

Proof [Proof of Theorem 3] For any time interval $\mathcal{I} = [s, t]$ with $|\mathcal{I}| \leq L$, if $|\mathcal{I}| \leq T/M$ then the regret bound holds trivially. Otherwise, there must be a round $s' \in \mathcal{I}$ such that $s' - s \leq T/M$, and there is a new copy of BISTRO+ added to the pool at round s'. Denote this new copy by $\mathcal{B}_{i^\star}$. The interval regret on $\mathcal{I}$ is then clearly bounded by T/M plus the interval regret on $[s', t]$.

Let $c_t(a) = 1 - r_t(a)$, $\forall a \in [K]$ and m_τ be the value of m at round τ before Line 15. Then for any policy π, we rewrite the interval regret on $[s', t]$ as:

$$
\mathbb{E}\left[\sum_{\tau=s'}^t r_\tau(\pi(x_\tau)) - r_\tau(a_\tau)\right] = \mathbb{E}\left[\sum_{\tau=s'}^t c_\tau(a_\tau) - \ell_\tau(i^\star) + \ell_\tau(i^\star) - c_\tau(\pi(x_\tau))\right]
$$

$$
= \mathbb{E}\left[\sum_{\tau=1}^t c_\tau(a_\tau) - \ell_\tau(i^\star)\right] + \mathbb{E}\left[\sum_{\tau=s'}^t c_\tau(a_\tau^{i^*}) - c_\tau(\pi(x_\tau))\right]
$$

$$
= \mathbb{E}\left[\sum_{\tau=1}^t \sum_{i=1}^{m_\tau} q_\tau(i)\ell_\tau(i) - \ell_\tau(i^\star)\right] + \mathbb{E}\left[\sum_{\tau=s'}^t c_\tau(a_\tau^{i^*}) - c_\tau(\pi(x_\tau))\right]
$$

$$
= \mathbb{E}\left[\sum_{\tau=1}^t \sum_{i=1}^M \bar{w}_\tau(i)\ell_\tau(i) - \ell_\tau(i^\star)\right] + \mathbb{E}\left[\sum_{\tau=s'}^t c_\tau(a_\tau^{i^*}) - c_\tau(\pi(x_\tau))\right]
$$

$$(*)$$

where the second equality uses the fact $c_\tau(a_\tau) = \ell_\tau(i^\star)$ for $\tau < s'$ and $\mathbb{E}_{i_\tau \sim q_\tau}[\ell_\tau(i^\star)] = c_\tau(a_\tau^{i^*})$ for $\tau \geq s'$, and the last equality holds because

$$
\sum_{i=1}^M \bar{w}_\tau(i)\ell_\tau(i) = \left(\sum_{i=1}^{m_\tau} \bar{w}_\tau(i)\right)\sum_{i=1}^{m_\tau} q_\tau(i)\ell_\tau(i) + \left(\sum_{i=m_\tau+1}^M \bar{w}_\tau(i)\right)\sum_{i=1}^{m_\tau} q_\tau(i)\ell_\tau(i) = \sum_{i=1}^{m_\tau} q_\tau(i)\ell_\tau(i).
$$

Here the first equality follows since $q_\tau(i) \left(\sum_{j=1}^{m_\tau} \bar{w}_\tau(j) \right) = \bar{w}_\tau(i)$ for $i \leq m_\tau$ and $\ell_\tau(i) = \sum_{j=1}^{m_\tau} q_\tau(j)\ell_\tau(j)$ for $i > m_\tau$ by definitions.

Next we bound the two terms in $(*)$. The first term is essentially the regret of the master, corresponding to the update in Line 10. Using results from (Wei and Luo, 2018),[6] we obtain

$$
\mathbb{E}\left[\sum_{\tau=1}^{t} \sum_{i=1}^{M} \bar{w}_\tau(i)\ell_\tau(i) - \ell_\tau(i^\star) \right]
$$

$$
\leq \widetilde{\mathcal{O}}\left(\frac{M}{\eta} + \eta \sum_{\tau=1}^{t} \mathbb{E}\left[w_\tau(i^\star)(\ell_\tau(i^\star) - (1 - r_\tau(a_\tau)))^2 \right] \right) - \mathbb{E}\left[\frac{\rho_{t,i^\star}}{40\eta \ln T} \right]
$$

$$
= \widetilde{\mathcal{O}}\left(\frac{M}{\eta} + \eta \sum_{\tau=s'}^{t} \mathbb{E}\left[w_\tau(i^\star)(\ell_\tau(i^\star) - (1 - r_\tau(a_\tau)))^2 \right] \right) - \mathbb{E}\left[\frac{\rho_{t,i^\star}}{40\eta \ln T} \right]
$$

$$
\leq \widetilde{\mathcal{O}}\left(\frac{M}{\eta} + L\eta \right) - \mathbb{E}\left[\frac{\rho_{t,i^\star}}{40\eta \ln T} \right]
$$

where the equality holds because by construction $\ell_\tau(i^\star) = (1 - r_\tau(a_\tau))$ for all $\tau < s'$.[7] For the second term in $(*)$, we apply Lemma 17 of (Agarwal et al., 2017) to obtain

$$
\mathbb{E}\left[\sum_{\tau=s'}^{t} c_\tau(a_\tau^{i^\star}) - c_\tau(\pi(x_\tau)) \right] = \mathbb{E}\left[\rho_{t,i^\star}^{1/3} \right] (LK)^{\frac{2}{3}} (\ln N)^{\frac{1}{3}}.
$$

Combining and proceeding similarly as the proof of Theorem 7 of (Agarwal et al., 2017) we have

$$
\mathbb{E}\left[\sum_{\tau=s'}^{t} r_\tau(\pi(x_\tau)) - r_\tau(a_\tau) \right] \leq \widetilde{\mathcal{O}}\left(\frac{M}{\eta} + L\eta \right) - \mathbb{E}\left[\frac{\rho_{t,i^\star}}{40\eta \ln T} \right] + \mathbb{E}\left[\rho_{t,i^\star}^{1/3} \right] (LK)^{\frac{2}{3}} (\ln N)^{\frac{1}{3}}
$$

$$
\leq \widetilde{\mathcal{O}}\left(\frac{M}{\eta} + L\eta + LK\sqrt{\eta \ln N} \right).
$$

Adding back the extra T/M term discussed above and plugging in the value of η and M lead to $\widetilde{\mathcal{O}}(T^{\frac{1}{4}}(LK)^{\frac{1}{2}}(\ln N)^{\frac{1}{4}} + \sqrt{T})$. Note that the term $\sqrt{T}$ is dominant only when $L \leq \sqrt{T}$, in which case even the first term is superlinear in L and becomes vacuous. We can therefore drop the second term and obtain the claimed bound. ∎

We finally include the dynamic regret guarantee for this algorithm, which is again a direct application of Lemma 6 combined with Theorem 3, similar to Corollary 2.

Corollary 5 *In the transductive setting, Algorithm 5 guarantees*

$$
\mathbb{E}\left[\sum_{t=1}^{T} r_t(\pi_t^\star(x_t)) - r_t(a_t) \right] = \widetilde{\mathcal{O}}\left(\min_{0 \leq L' \leq L} \left\{ \frac{T}{L'} \left(T^{\frac{1}{4}}(LK)^{\frac{1}{2}}(\ln N)^{\frac{1}{4}} \right) + \Delta L' \right\} \right).
$$

If Δ is known, optimally setting $L = \min\{ T^{\frac{5}{6}} K^{\frac{1}{3}} (\ln N)^{\frac{1}{6}} / \Delta^{\frac{2}{3}}, T \}$ gives $\widetilde{\mathcal{O}}(\Delta^{\frac{1}{3}} T^{\frac{5}{6}} K^{\frac{1}{3}} (\ln N)^{\frac{1}{6}} + T^{\frac{3}{4}} K^{\frac{1}{2}} (\ln N)^{\frac{1}{4}})$; otherwise, setting $L = T^{\frac{5}{6}}$ gives $\widetilde{\mathcal{O}}((\sqrt{\Delta} + 1)T^{\frac{5}{6}} K^{\frac{1}{2}} (\ln N)^{\frac{1}{4}})$.

---

6. This is not explicitly given in (Wei and Luo, 2018), but is a direct application of their Theorem 2 with $m_{t,i} = \ell_{t,i_t}$ in their notation.

7. This is the exact place where we obtain some improvement over the original CORRAL by using results of (Wei and Luo, 2018).

Proof sketch. The proof follows the same procedure as in Corollary 2: partition $[1, T]$ into $\frac{T}{L'}$ intervals each of length L', plug in the interval regret guarantee for each interval (Theorem 3), and then apply Lemma 6 to obtain the claimed dynamic regret. ∎

Appendix H. Interval Regret for ADA-BINGREEDY

Theorem 28 *Let $\mathcal{I}$ be an interval with $\Delta_{\mathcal{I}} = 0$. Then* ADA-BINGREEDY *with parameter δ guarantees that with probability at least $1 - 5\delta$,*

$$\sum_{t \in \mathcal{I}} r_t(\pi(x_t)) - r_t(a_t) = \widetilde{\mathcal{O}}(\sqrt{K'}T^{\frac{3}{4}} + K'\sqrt{T})$$

for any $\pi \in \Pi$, where $K' = K \ln(N/\delta)$.

Proof In the proof of Lemma 21, we have shown that with probability at least $1 - \delta/2$, if there is no distribution change, an epoch will not rerun. This implies that with probability $1 - \delta/2$, the rerun is triggered at most once in $\mathcal{I}$. Below we assume this event indeed holds. Let $\mathcal{I} = \mathcal{I}_1 \cup \{t'\} \cup \mathcal{I}_2$, where in $\mathcal{I}_1$ and $\mathcal{I}_2$ rerun is never triggered.

We can view $\mathcal{I}_2$ as a fresh epoch with no distribution change in it. Reusing Lemma 23, 24, and 25's *epoch regret* intermediate results (i.e., those **Regret in an epoch** paragraphs in the proofs) with $S_{\mathcal{I}_2} = 1$, we get

$$\sum_{t \in \mathcal{I}_2} \mathbb{E}_t[r_t(\pi(x_t)) - r_t(a_t)] = \widetilde{\mathcal{O}}(\sqrt{K'}T^{\frac{3}{4}} + K'\sqrt{T})$$

with probability at least $1 - 2\delta$.

For $\mathcal{I}_1$, we can decompose it into $\mathcal{J}_{j'} \cup \mathcal{J}_{j'+1} \cup \cdots \cup \mathcal{J}_{j^*}$, where $\mathcal{J}_{j'+1}, \cdots, \mathcal{J}_{j^*-1}$ are complete blocks with block indices $j' + 1, \ldots, j^* - 1$ respectively, while $\mathcal{J}_{j'}$ and $\mathcal{J}_{j^*}$ are possibly incomplete blocks (rerun is triggered in $\mathcal{J}_{j^*}$). For $j = j' + 1, \ldots, j^*$, we can bound the regret in flat exploitation bins in $\mathcal{J}_j$ by reusing the *block regret* result in the proof of Lemma 23 with $S_{\mathcal{J}_j} = 1$. Applying the last bound in the **Regret in a block** part of Lemma 23, the regret in flat exploitation bins in $\mathcal{J}_{j'+1} \cup \ldots \cup \mathcal{J}_{j^*}$ can be bounded by

$$\sum_{t \in \mathcal{J}_{j'+1} \cup \cdots \cup \mathcal{J}_{j^*}} \mathbb{E}_t[r_t(\pi(x_t)) - r_t(a_t)]\mathbf{1}\{t \text{ is in flat-exploitation bins}\}$$

$$= \widetilde{\mathcal{O}}\left(\sum_{j=j'+1}^{j^*} \sqrt{K'}\left(2^{(j-1) \times \frac{3}{4}}\right) + K'\left(2^{(j-1) \times \frac{1}{2}}\right)\right) + \widetilde{\mathcal{O}}(\sqrt{K'}T^{\frac{2}{3}} + K'T^{\frac{1}{3}})$$

$$= \widetilde{\mathcal{O}}(\sqrt{K'}T^{\frac{3}{4}} + K'\sqrt{T}).$$

The sum of regret in exploration bins or in non-flat bins can be bounded by $\widetilde{\mathcal{O}}(T^{\frac{3}{4}})$ with probability $1 - 2\delta$ by Lemma 24 and 25's epoch regret results. Finally, the regret in $\mathcal{J}_{j'}$ can be bounded by $|\mathcal{J}_{j'}| = \widetilde{\mathcal{O}}(\sqrt{T})$. Combining all above and using Hoeffding-Azuma inequality complete the proof. ∎