Tracking Most Significant Shifts in Nonparametric Contextual Bandits

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We study nonparametric contextual bandits where Lipschitz mean reward functions may change over time. We first establish the minimax dynamic regret rate in this less understood setting in terms of number of changes L and total-variation V, both capturing all changes in distribution over context space, and argue that state-of-the-art procedures are suboptimal in this setting.

Next, we tend to the question of an *adaptivity* for this setting, i.e. achieving the minimax rate without knowledge of L or V. Quite importantly, we posit that the bandit problem, viewed local at a given context X_t, should not be affected by reward changes in other parts of context space $\mathcal{X}$. We therefore propose a notion of *change* that better accounts for locality, and thus counts significantly less changes than L and V. Our main result is to show that this more strict notion of change, which we term *experienced significant shifts*, can in fact be adapted to. As in previous work on non-stationary MAB (Suk and Kpotufe, 2022), not only do our results capture changes only at the experienced contexts x, but also only the most *significant* in terms of changes in mean rewards (e.g., only count severe best-arm changes at x).

1 Introduction

Contextual bandits model sequential decision making problems where the reward of a chosen action depends on an observed context X_t at time t, e.g., a consumer's profile, a medical patient's history. The goal is to maximize the total rewards over time of chosen actions, as informed by seen contexts. As such, one suitable measure of performance is that of *dynamic regret*, which compares earned rewards to a time-varying oracle maximizing mean rewards at X_t. While it is often assumed in the bulk of works in this setting that rewards distributions remain stationary over time, it is understood that in practice, environmental changes induce nontrivial changes in rewards.

In fact, the problem of non-stationary environments has received a surge of attention in the simpler non-contextual Multi-Arm-Bandits (MAB) setting, while the more challenging contextual case remains ill-understood. In particular in the contextual case, some recent works of Wu et al. [2018], Luo et al. [2018], Chen et al. [2019], Wei and Luo [2021] consider *parametric settings*, i.e. where reward functions belong to fixed parametric family, and show that one may achieve rates adaptive to an unknown number of L of shifts in rewards or to a notion of total-variation V, both acccounting for all changes over time and context space. Instead here, we consider a much larger class of reward functions, namely Lipschitz rewards, corresponding to the natural assumption that closeby contexts have similar rewards even as reward distributions change.

As a first result for this nonparametric setting, we establish some minimax lower-bounds as a baseline in terms of either L or V, and argue that state-of-the-art procedures for the parametric case—extended to the class of Lipschitz functions—do not achieve these baselines.

We then turn attention to whether such baselines may be achieved *adaptively*, i.e., without knowledge of L or V. The answer as we show is affirmative, and more importantly, some much weaker notions of change may be adapted to; for intuition, while L or V accounts for any change at any time over the context space (say $\mathcal{X}$), it may be that all changes are relegated to parts of the space irrelevant to observed contexts X_t at the time they are played. For instance, suppose at time t, we observe $X_t = x_0$, then it may not make sense to count changes that happen at some other x_1 far from x_0, or changes that happened at x_0 itself but far back in time.

We therefore propose a new parameterization of change, termed *experienced significant shifts* that better accounts for the locality of changes in time and space, and as such may register much less changes than either L or V. As a sanity check, we show that an oracle policy which restarts only at experienced significant shifts can attain enhanced regret rates in terms of the number $\tilde{L} = \tilde{L}(X_1, \ldots, X_T)$ of such experienced shifts (Proposition 2), a rate always no worse that the baseline we first established in terms of L and V.

Our main result is to show that *experienced significant shifts* can be adapted to (Theorem 3), i.e., with no prior knowledge of such shifts. Importantly, the result holds in both stochastic environments, and in (oblivious) adversarial ones with no change to our notion, algorithmic approach, nor analysis. Furthermore, similar to recent advances in the non-contextual case [Abbasi-Yadkori et al., 2022, Suk and Kpotufe, 2022], an *experienced shift* is only triggered under *severe changes* such as changes of best arms locally at a context X_t. An added difficulty in the contextual case is that we cannot hope to observe rewards for a given arm (action) repeatedly at X_t as the context may only appear once, and have to rely on carefuly chosen nearby points to identify unknown shifts in reward at X_t.

1.1 Other Related Work

Nonparametric Contextual Bandits. The stationary bandits with covariates (where rewards and contexts follow a joint distribution) was first introduced in a one-armed bandit problem [Woodroofe, 1979, Sarkar, 1991], with the nonparametric model first studied by Yang et al. [2002]. Minimax regret rates, based on a margin condition, were first established for the two-armed bandit in Rigollet and Zeevi [2010] and generalized to any finite number of arms in Perchet and Rigollet [2013], with further insights thereafter [Qian and Yang, 2016a,b, Reeve et al., 2018, Guan and Jiang, 2018, Gur et al., 2022, Hu et al., 2020, Arya and Yang, 2020, Suk and Kpotufe, 2021, Cai et al., 2022]. However, the mentioned works all assume a stationary distribution of rewards over contexts. Blanchard et al. [2023] studies non-stationary nonparametric contextual bandits, but in the much-different context of universal learning, concerning when sublinear regret is achievable asymptotically.

Lipschitz contextual bandits appears as part of studies on broader infinite-armed settings [Lu et al., 2009, Krishnamurthy et al., 2019]. Related, Slivkins [2014] allows for non-stationary (i.e., obliviously adversarial) environments, but only studies regret to the (per-context) best arm in hindsight.

Realizable contextual bandits posits that the regression function capturing mean rewards in contexts lies in some known class of regressors $\mathcal{F}$, over which one can do empirical risk minimization [Foster et al., 2018, Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2021]. While this setting can recover Lipschitz contextual bandits, the only result on non-stationary guarantees to our knowledge is Wei and Luo [2021], which yields suboptimal dynamic regret (see Table 1).

Non-Stationary Bandits and RL. In the simpler non-contextual bandits, changing reward distributions (a.k.a. *switching bandits*) was introduced in Garivier and Moulines [2011] and further explored with various assumptions and formulations [Besbes et al., 2019, Karnin and Anava, 2016, Allesiardo et al., 2017, Liu et al., 2018, Wei and Srivatsva, 2018, Besson et al., 2022, Cao et al., 2019, Mukherjee and Maillard, 2019]. While these earlier works focused on algorithmic design assuming knowledge of non-stationarity, such a strong assumption was removed via the *adaptive* procedures of Auer et al. [2019], Chen et al. [2019]. In followup works, Abbasi-Yadkori et al. [2022], Suk and Kpotufe [2022] show that tighter dynamic regret rates are possible, scaling only with severe changes in best arm.

The ideas from non-stationary MAB were extended to various contextual bandit settings by Wu et al. [2018] (for linear mean rewards in contexts), Luo et al. [2018], Chen et al. [2019] (for finite policy classes), and Wei and Luo [2021] (for realizable mean reward functions).

There have also been extensions of these ideas to various reinforcement learning setups [Jaksch et al., 2010, Gajane et al., 2018, Ortner et al., 2020, Cheung et al., 2020, Fei et al., 2020, Mao et al., 2021, Zhou et al., 2022, Touati and Vincent, 2020, Domingues et al., 2021, Chi Cheung et al., 2019, Domingues et al., 2021, Ding and Lavaei, 2023, Wei and Luo, 2021, Lykouris et al., 2021, Wei et al., 2022, Chen and Luo, 2022]. Among these works, only Domingues et al. [2021] can recover Lipschitz contextaul bandits, whereupon we find their dynamic regret bounds are suboptimal (see Table 1).

Again, the typical aim of aforementioned works on contextual bandits or RL is to minimize a notion of dynamic regret in terms of the number of changes L or total-variation V. As such, regardless of setting, known guarantees in said works do not involve tighter notions of experienced non-stationarity.

2 Problem Formulation

2.1 Contextual Bandits with Changing Rewards

Preliminaries. We assume a finite set of arms $[K] \doteq \{1, 2 \ldots, K\}$. Let $Y_t \in [0,1]^K$ denote the vector of rewards for arms $a \in [K]$ at round $t \in [T]$ (horizon T), and X_t the observed context at that round, lying in $\mathcal{X} \doteq [0,1]^d$, which have joint distribution $(X_t, Y_t) \sim \mathcal{D}_t$. We let $\mathbf{X}_t \doteq \{X_s\}_{s \leq t}, \mathbf{Y}_t \doteq \{Y_s\}_{s \leq t}$ denote the observed contexts and (observed and unobserved) rewards from rounds 1 to t. In our setting, an oblivious adversary decides a sequence of (independent) distributions on $\{(X_t, Y_t)\}_{t \in [T]}$ before play.

Notation. *The* **reward function** $f_t : \mathcal{X} \to [0,1]^K$ *is* $f_t^a(x) \doteq \mathbb{E}[Y_t^a | X_t = x]$, $a \in [K]$, *and captures the mean rewards of arm a at context x and time t.*

A *policy* chooses actions at each round t, based on observed contexts (up to round t) and passed rewards, whereby at each round t only the reward Y_t^a of the chosen action a is revealed. Formally:

Definition 1 (Policy). *A policy $\pi \doteq \{\pi_t\}_{t \in \mathbb{N}}$ is a random sequence of functions $\pi_t : \mathcal{X}^t \times [K]^{t-1} \times [0,1]^{t-1} \to [K]$. In the case of a* **randomized** *policy, i.e., where π_t in fact maps to distributions on $[K]$, In an abuse of notation, in the context of a sequence of observations till round t, we'll let $\pi_t \in [K]$ denote the (possibly random) action chosen at round t.*

The performance of a policy is evaluated using the *dynamic regret*, defined as follows:

Definition 2. *Fix a context sequence $\boldsymbol{X}_T$. Define the* **dynamic regret** *of a policy π, as*

$$R_T(\pi, \boldsymbol{X}_T) \doteq \sum_{t=1}^{T} \max_{a \in [K]} f_t^a(X_t) - f_t^{\pi_t}(X_t).$$

Thus, we seek a policy π that minimizes $\mathbb{E}[R_T(\pi, \mathbf{X}_T)]$ where the expectation is over $\mathbf{X}_T$, $\mathbf{Y}_T$, and any randomness in π.

Notation. *As much of our analysis focuses on the gaps in mean rewards between arms at observed contexts X_t, the following notation will serve useful. Let $\delta_t(a', a) \doteq f_t^{a'}(X_t) - f_t^a(X_t)$ denote the* **relative gap** *of arms a to a' at round t at context X_t. Define the* **worst gap** *of arm a as $\delta_t(a) \doteq \max_{a' \in [K]} \delta_t(a', a)$, corresponding to the instantaneous regret of playing a at round t and context X_t. Thus, the dynamic regret can be written as $\sum_{t \in [T]} \mathbb{E}[\delta_t(\pi_t)]$. Additionally, let $\delta_t^{a',a}(x) \doteq f_t^{a'}(x) - f_t^a(x)$ and $\delta_t^a(x) \doteq \max_{a' \in [K]} \delta_t^{a',a}(x)$ be the* **gap functions** *mapping $\mathcal{X} \to [0,1]$.*

2.2 Nonparametric Setting

We assume, as in prior work on nonparametric contextual bandits [Rigollet and Zeevi, 2010, Perchet and Rigollet, 2013, Slivkins, 2014, Reeve et al., 2018, Guan and Jiang, 2018, Suk and Kpotufe, 2021], that the reward function is 1-Lipschitz.

Assumption 1 (Lipschitz f_t). *For all rounds $t \in \mathbb{N}$, $a \in [K]$ and $x, x' \in \mathcal{X}$,*

$$|f_t^a(x) - f_t^a(x')| \leq \|x - x'\|_\infty. \tag{1}$$

For ease of presentation, we assume the contextual marginal distribution μ_X remains the same across rounds. Furthermore, we make a standard *strong density assumption* on μ_X, which is typical in this nonparametric setting [Audibert and Tsybakov, 2007, Perchet and Rigollet, 2013, Qian and Yang, 2016a,b, Gur et al., 2022, Hu et al., 2020, Arya and Yang, 2020, Cai et al., 2022]. This holds, e.g. if μ_X has a continuous Lebesgue density on $[0,1]^d$, and ensures good coverage of the context space.

Assumption 2 (Strong Density Condition)**.** *There exist $C_d, c_d > 0$ s.t. $\forall \ell_\infty$ balls $B \subset [0,1]^d$ of diameter $r \in (0,1]$:*

$$C_d \cdot r^d \geq \mu_X(B) \geq c_d \cdot r^d. \tag{2}$$

Remark 1. *We can in fact relax the above assumptions on context marginals so that $\mu_{X,t}(\cdot)$ is changing with time t and the above strong density assumption is satisfied with different constants $C_{d,t}, c_{d,t}$. Our procedures in the end will not require knowledge of any $C_{d,t}, c_{d,t}$.*

2.3 Model Selection

A common algorithmic approach in nonparametric contextual bandits, starting from earlier work [Rigollet and Zeevi, 2010, Perchet and Rigollet, 2013], is to discretize or partition the context space $\mathcal{X}$ into *bins* where we can maintain local reward estimates. These bins have a natural hierarchical tree structure which we first elaborate.

Definition 3 (Partition Tree)**.** *Let $\mathcal{R} \doteq \{2^{-i} : i \in \mathbb{N} \cup \{0\}\}$, and let $\mathcal{T}_r, r \in \mathcal{R}$ denote a regular partition of $[0,1]^d$ into hypercubes (which we refer to as* **bins***) of side length (a.k.a. bin size) r. We then define the dyadic* **tree** *$\mathcal{T} \doteq \{\mathcal{T}_r\}_{r\in\mathcal{R}}$, i.e., a hierarchy of nested partitions of $[0,1]^d$. We will refer to the* **level** *r* **of** *T as the collection of bins in partition $\mathcal{T}_r$. The* **parent** *of a bin $B \in \mathcal{T}_r, r < 1$ is the bin $B' \in \mathcal{T}_{2r}$ containing B;* **child**, **ancestor** *and* **descendant** *relations follow naturally. The notation $T_r(x)$ will then refer to the bin at level r containing x.*

Note that, while in the above definition, $\mathcal{T}$ has infinite levels $r \in \mathcal{R}$, at any round t in a procedure, we implicitly only operate on the subset of $\mathcal{T}$ containing data.

Key in securing good regret is then finding the optimal level $r \in \mathcal{R}$ of discretization (balancing regression bias and variance), which over n stationary rounds is known to be $\propto (K/n)^{\frac{1}{2+d}}$ [Rigollet and Zeevi, 2010]. We introduce the following general notation, useful later in the approaching the non-stationary problem, for associating the size of a level to an intervals of rounds.

Notation 1 (Level)**.** *For $n \in \mathbb{N} \cup \{0\}$, let r_n be the largest $2^{-m} \in \mathcal{R}$ such that $(K/n)^{\frac{1}{2+d}} \geq 2^{-m}$.*

We use $\mathcal{T}_m, T_m(x)$ as shorthand to denote (respectively) the tree $\mathcal{T}_r$ of level $r = r_m$ and the (unique) bin at level r_m containing x.

3 Results Overview

3.1 Minimax Lower Bounds Under Global Shifts

As a baseline, we start with some basic lower-bounds under the simplest parametrizations of changes in rewards which have appeared in the literature, namely a *global number of shifts*, and *total variation*.

Definition 4 (Global Number of Shifts)**.** *Let $L \doteq \sum_{t=2}^{T} \mathbf{1}\{\exists x \in \mathcal{X}, a \in [K] : f_t^a(x) \neq f_{t-1}^a(x)\}$ be the number of* global shifts, *i.e., it counts every change in mean-reward overtime and over $\mathcal{X}$ space.*

Definition 5 (Total Variation)**.** *Define $V_T \doteq \sum_{t=2}^{T} \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{TV}$ where recall $\mathcal{D}_t \in \mathcal{X} \times [0,1]^K$ is the joint distribution on context and rewards at time t.*

We have the following initial result (for two-armed bandits) to serve as baseline for this study.

Theorem 1 (Dynamic Regret Lower Bound)**.** *Suppose there are $K = 2$ arms. For $V, L \in [0,T]$, let $\mathcal{P}(V,L,T)$ be the family of joint distributions $\mathcal{D} \doteq \{\mathcal{D}_t\}_{t\in[T]}$ with either total variation $V_T \leq V$ or at most L global shifts. Then, there exists a constant $c > 0$ such that:*

$$\sup_{\mathcal{D}\in\mathcal{P}(V,L,T)} \mathbb{E}_{\mathcal{D}}[R(\pi, \boldsymbol{X}_T)] \geq c\left(T^{\frac{1+d}{2+d}} + T^{\frac{2+d}{3+d}} \cdot V^{\frac{1}{3+d}}\right) \wedge \left((L+1)^{\frac{1}{2+d}} T^{\frac{1+d}{2+d}}\right). \tag{3}$$

Remark 2. *Note setting $d = 0$ in Theorem 1 recovers the established non-contextual minimax rate of $(\sqrt{T} + T^{2/3}V_T^{1/3}) \wedge \sqrt{(L+1)\cdot T}$.*

Achievability of Miminimax Lower-Bound (3). We are interested in whether the rates of (3) are achievable, with, or without knowledge of relevant parameters. First, we note that no existing algorithm currently guarantees a rate that matches (3). See Table 1 for a rate comparison (details in Appendix A).

In particular, the prior adaptive works [Chen et al., 2019, Wei and Luo, 2021] both rely on the approach of randomly scheduling *replays* of stationary algorithms to detect unknown non-stationarity. However, the scheduling rate is designed to safeguard against their parametric $\sqrt{LT} \wedge V_T^{1/3} T^{2/3}$ regret rates and thus lead to suboptimal dependence on L and V_T.

However, a simple back of the envelope calculation indicates that the rate in (3) may be attainable, at least given some distributional knowledge: a procedure restarting at each shift will incur regret, over L equally spaced shifts, $(L+1) \cdot \left(\frac{T}{L+1}\right)^{\frac{1+d}{2+d}} \approx L^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}$.

As it turns out as we will show in the next section, (3) is indeed attainable, even adaptively; in fact, this is shown via a more optimistic problem parametrization as described next.

	Dynamic Regret Upper Bound
ADA-ILTCB [Chen et al., 2019]	$\left(L^{1/2} \cdot T^{\frac{1+d}{2+d}}\right) \wedge \left(V_T^{1/3} \cdot T^{\frac{2+d}{3+d} + \frac{d}{3(2+d)(3+d)}}\right)$
MASTER with FALCON [Wei and Luo, 2021]	$\left(L^{1/2} \cdot T^{\frac{1+d}{2+d}}\right) \wedge \left(V_T^{1/3} \cdot T^{\frac{2+d}{3+d} + \frac{d}{3(2+d)(3+d)}}\right)$
KeRNS [Domingues et al., 2021] (non-adaptive)	$V_T^{1/3} T^{\frac{2+d}{3+d} + O(1/d)}$
Minimax Lower-Bound	$\left(L^{\frac{1}{2+d}} T^{\frac{1+d}{2+d}}\right) \wedge \left(V_T^{\frac{1}{3+d}} T^{\frac{2+d}{3+d}}\right)$

Table 1: Existing dynamic Regret Upper-Bounds appear suboptimal in the Lipschitz setting.

3.2 A New Problem Parametrization: Experienced Significant Shifts.

As discussed in Section 1, typical approaches in our setting discretize the context space $\mathcal{X}$ into bins, each of which is treated as an MAB instance. At a high level, our new measure of non-stationarity will trigger an **experienced significant shift** when the observed context X_t arrives in a bin $B \in \mathcal{T}$ where there has been a severe change in local best arm, *w.r.t. the observed data in that bin*.

We first define a notion of **significant regret** for an arm $a \in [K]$ *locally* within a bin $B \in \mathcal{T}$. We say arm a incurs **significant regret** in bin B on interval I if:

$$\sum_{s \in I} \delta_s(a) \cdot \mathbf{1}\{X_s \in B\} \geq \sqrt{K \cdot n_B(I)} + r(B) \cdot n_B(I), \tag{$\star$}$$

where $n_B(I) \doteq \sum_{s \in I} \mathbf{1}\{X_s \in I\}$. The intuition for ($\star$) is as follows: suppose that, over n separate rounds, we observe the same context $X_s = x_0$ in bin B. Then, arm a would be considered unsafe in the local bandit problem at context x_0 if its regret exceeds $\sqrt{K \cdot n}$ (i.e., the first term on the above RHS), which is a safe regret to pay for the non-contextual problem. Our broader notion ($\star$) extends this over the bin B by also accounting for the bias (i.e., the second term on the above RHS) of observing X_s near a given context $x_0 \in B$.

We then propose to record an *experienced significant shift* when we experience a context X_t, for which there is no safe arm to play in the sense of ($\star$).

Definition 6. *Fix the context sequence $X_1, X_2, \ldots, X_T$.*

• *We say an arm $a \in [K]$ is* **unsafe at context** *$x \in \mathcal{X}$* **on** *I if there exists a bin $B \in \mathcal{T}$ containing x such that arm a incurs significant regret ($\star$) in bin B on I.*

We then have the following recursive definition:

• *Let $\tau_0 = 1$. We then have the following recursive definition: define the $(i+1)$-th* **experienced significant shift** *as the earliest time $\tau_{i+1} \in (\tau_i, T]$ such that every arm $a \in [K]$ is unsafe at X_t on some interval $I \subset [\tau_i, \tau_{i+1}]$. We refer to intervals $[\tau_i, \tau_{i+1}), i \geq 0$, as* **experienced significant phases***. The unknown number of such phases (by time T) is denoted $\tilde{L}$, whereby $[\tau_{\tilde{L}-1}\tau_{\tilde{L}})$, for $\tau_{\tilde{L}} \doteq T+1$, is the last phase.*

Remark 3 (Significant Shifts Depend on Contexts)**.** *It should be understood that the significant shifts τ_i and $\tilde{L}$ depend on $\mathbf{X}_T$ and mean rewards $\{f_t^a(X_t)\}_{t \in [T], a \in [K]}$, but not the realized rewards $\mathbf{Y}_T$. For simplicity of presentation, we will not make the dependence on $\mathbf{X}_T$ explicit in most places where $\tau_i, \tilde{L}$ are mentioned.*

It's clear from Definition 6 and ($\star$) that only changes in the mean rewards $f_t^a(x)$ at experienced contexts $x \in \mathbf{X}_T$ are counted, and that they are only counted when experienced. Furthermore, an experienced significant shift τ_i implies a best-arm change at X_{τ_i} since, by smoothness (Assumption 1), and ($\star$) we have

$$\sum_{s \in I} \delta_s^a(X_{\tau_i}) \cdot \mathbf{1}\{X_s \in B\} \geq \sum_{s \in I} \delta_s(a) \cdot \mathbf{1}\{X_s \in B\} - r(B) \sum_{s \in I} \mathbf{1}\{X_s \in B\} > 0.$$

Thus, $\tilde{L} \leq L + 1$, the global count of shifts.

On the other hand, so long as an experienced significant shift does not occur, there will be arms safe to play at each context X_t. As a result, procedures need not restart exploration so long as unsafe arms can be quickly ruled out.

As a warmup to presenting our main regret bounds and algorithms, we'll first consider an oracle procedure which restarts only at experienced significant shifts.

Definition 7 (Oracle Procedure)**.** *For each round t in phase $[\tau_i, \tau_{i+1})$, define a good arm set $\mathcal{G}_t$ as the set of **safe** arms, i.e., arms which do not yet satisfy ($\star$) in bin $T_r(X_t)$ for $r = r_{\tau_{i+1} - \tau_i}$ (recall from Subsection 2.3 that this is the oracle choice of level over phase $[\tau_i, \tau_{i+1})$).*

*Then, define an **oracle procedure** π: at each round t, π plays a random arm $a \in \mathcal{G}_t$ w.p. $1/|\mathcal{G}_t|$.*

We then claim such an oracle procedure attains an enhanced dynamic regret rate in terms of the significant shifts $\{\tau_i\}_i$ which recovers the minimax lower bound in terms of global number of shifts L and total variation V_T from before.

Proposition 2 (Sanity Check)**.** *We have the oracle procedure π of Definition 7 satisfies with probability at least $1 - 1/T^2$ w.r.t. the randomness of $\boldsymbol{X}_T$: for some $C > 0$*

$$\mathbb{E}_\pi[R_T(\pi, \boldsymbol{X}_T) \mid \boldsymbol{X}_T] \leq C \log(K) \log(T) \sum_{i=1}^{\tilde{L}(\boldsymbol{X}_T)} (\tau_i(\boldsymbol{X}_T) - \tau_{i-1}(\boldsymbol{X}_T))^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}}.$$

Proof. See Appendix C. $\square$

By Jensen's inequality on the concave function $z \mapsto z^{\frac{1+d}{2+d}}$, the above regret rate is at most $\tilde{L}(\mathbf{X}_T)^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}} \ll L^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}$. At the same time, the rate is also faster than $V_T^{\frac{1}{3+d}} T^{\frac{1+d}{2+d}}$ (see Corollary 5). We next aim to design an algorithm which can attain the same regret without knowledge of τ_i or $\tilde{L}$.

3.3 Main Results: Adaptive Upper-bounds

Our main result is a dynamic regret upper bound of similar order to Proposition 2 *without knowledge of the environment, e.g., the significant shift times, or the number of significant phases*. It is stated for our algorithm CMETA (Algorithm 1 of Section 4), which, for simplicity, requires knowledge of the time horizon T (knowledge of T removable using doubling tricks).

Theorem 3. *Let π denote the* CMETA *procedure. Let $\{\tau_i(\boldsymbol{X}_T)\}_{i=0}^{\tilde{L}+1}$ denote the unknown experienced significant shifts (Definition 6). We then have with probability at least $1 - 1/T^2$ w.r.t. the randomness of $\boldsymbol{X}_T$, for some $C > 0$:*

$$\mathbb{E}[R_T(\pi, \boldsymbol{X}_T) \mid \boldsymbol{X}_T] \leq C \log^4(T) \sum_{i=1}^{\tilde{L}(\boldsymbol{X}_T)} (\tau_i(\boldsymbol{X}_T) - \tau_{i-1}(\boldsymbol{X}_T))^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}}.$$

By Jensen's inequality, since the function $z \mapsto z^{\frac{1+d}{2+d}}$ is concave, the above regret rate is upper bounded by $\tilde{L}(\mathbf{X}_T)^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}}$,

Corollary 4 (Adapting to Experienced Significant Shifts)**.** *Under the conditions of Theorem 3, with probability at least $1 - 1/T^2$ w.r.t. the randomness in $\boldsymbol{X}_T$:*

$$\mathbb{E}[R_T(\pi, \boldsymbol{X}_T) \mid \boldsymbol{X}_T] \leq C \log^4(T) \cdot (K \cdot \tilde{L}(\boldsymbol{X}_T))^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}.$$

Note, this is tighter than the earlier mentioned $(K \cdot L)^{\frac{1}{2+d}} T^{\frac{1+d}{2+d}}$ rate. The next corollary asserts that Theorem 3 also recovers the optimal rate in terms of total-variation V_T.

Corollary 5 (Adapting to Total Variation)**.** *Under the conditions of Theorem 3, taking expectation over $\boldsymbol{X}_T$:*

$$\mathbb{E}[R_T(\pi, \boldsymbol{X}_T)] \leq C \log^4(T) \left(T^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}} + (V_T \cdot K)^{\frac{1}{3+d}} \cdot T^{\frac{2+d}{3+d}} \right).$$

4 Algorithm

Algorithm 1: **C**ontextual **M**eta-**E**limination while **T**racking **A**rms (CMETA)

Input: horizon T, set of arms $[K]$, tree $\mathcal{T}$ with levels $r \in \mathcal{R}$.

1 **Initialize:** round count $t \leftarrow 1$.

2 **Episode Initialization (setting global variables)**:

3 $\quad t_\ell \leftarrow t$. ; // t_ℓ indicates start of ℓ-th episode.

4 $\quad$For each bin $B \in \mathcal{T}$, set $\mathcal{A}_{\text{master}}(B) \leftarrow [K]$; // Initialize master candidate arm sets

5 $\quad$For each $m = 2, 4, \ldots, 2^{\lceil \log(T) \rceil}$ and $s = t_\ell + 1, \ldots, T$:

6 $\quad\quad$Sample and store $Z_{m,s} \sim \text{Bernoulli}\left(\left(\frac{1}{m}\right)^{\frac{1}{2+d}} \cdot \left(\frac{1}{s - t_\ell}\right)^{\frac{1+d}{2+d}} \right)$. ; // Set replay schedule.

7 Run Base-Alg $(t_\ell, T + 1 - t_\ell)$.

8 **if** $t < T$ **then** restart from Line 2 (i.e. start a new episode). ;

Algorithm 2: Base-Alg (t_{start}, m_0): binned successive elimination with randomized arm-pulls

Input: starting round t_{start}, scheduled duration m_0.

1 **Initialize**: $t \leftarrow t_{\text{start}}$ For each bin B at any level in $\mathcal{T}$, set $\mathcal{A}(B) \leftarrow [K]$

2 **while** $t \leq t_{\text{start}} + m_0$ **do**

3 $\quad$**Choose level in $\mathcal{R}$:** $r \leftarrow r_{t - t_{\text{start}}}$.

4 $\quad$Let $\mathcal{A}_t \leftarrow \mathcal{A}(B)$ and let $B \leftarrow T_r(X_t)$.

5 $\quad$Play a random arm $a \in \mathcal{A}_t$ selected with probability $1/|\mathcal{A}_t|$.

6 $\quad$Increment $t \leftarrow t + 1$.

7 $\quad$**if** $\exists m$ such that $Z_{m,t} > 0$ **then**

8 $\quad\quad$Let $m \doteq \max\{m \in \{2, 4, \ldots, 2^{\lceil \log(T) \rceil}\} : Z_{m,t} > 0\}$. ; // Set maximum replay length.

9 $\quad\quad$Run Base-Alg (t, m). ; // Replay interrupts.

10 $\quad$**Evict bad arms in bin B:**

11 $\quad\quad\mathcal{A}(B) \leftarrow \mathcal{A}(B) \backslash \{a \in [K] :$
$\quad\quad\quad\exists$ rounds $[s_1, s_2] \subseteq [t_{\text{start}}, t)$ s.t. (5) holds for bin $T_{s_2 - s_1}(X_t)\}$.

12 $\quad\quad\mathcal{A}_{\text{master}}(B) \leftarrow \mathcal{A}_{\text{master}}(B) \backslash \{a \in [K] :$
$\quad\quad\quad\exists$ rounds $[s_1, s_2] \subseteq [t_\ell, t)$ s.t. (5) holds for bin $T_{s_2 - s_1}(X_t)\}$.

13 $\quad\quad$**Refine candidate arms:** ; // Discard arms previously discarded in ancestor bins

14 $\quad\quad\quad\mathcal{A}(B) \leftarrow \cap_{B' \in \mathcal{T}, B \subseteq B'} \mathcal{A}(B')$.

15 $\quad\quad\quad\mathcal{A}_{\text{master}}(B) \leftarrow \cap_{B' \in \mathcal{T}, B \subseteq B'} \mathcal{A}_{\text{master}}(B')$.

16 $\quad$**Restart criterion: if** $\mathcal{A}_{\text{master}}(B) = \emptyset$ for some bin B **then** RETURN.;

17 RETURN.

We take a similar algorithmic approach to Suk and Kpotufe [2022], with several important modifications for our setting. The high-level strategy is to schedule multiple copies of a *base algorithm* (Algorithm 2) Base-Alg at random times and durations, in order to ensure updated and reliable estimation of the gaps in ($\star$). This allows fast enough detection of unknown experienced significant shifts.

Overview of Algorithm Hierarchy. Our main algorithm CMETA (Algorithm 1) proceeds in episodes, each of which begins by playing according to an initially scheduled base algorithm of possible duration equal to the number of rounds left till T. Base algorithms occasionally activate their own base algorithms of varying durations (Line 9 of Algorithm 2), called *replays*, according to a random schedule (stored in the variable $\{Z_{m,s}\}$). We refer to the base algorithm playing at round t as the *active base algorithm*. This induces a hierarchy of base algorithms, from *parent* to *child* instances of Base-Alg .

Choice of Level. Focusing on a single base algorithm now, each Base-Alg manages its own discretization of the context space $\mathcal{X} = [0,1]^d$, corresponding to a level $r \in \mathcal{R}$ (see Definition 3). Within each bin $B \in \mathcal{T}_r$ at the level r, candidate arms, maintained in a set $\mathcal{A}(B)$, are evicted according to estimates (4) of local gaps.

As said earlier in Subsection 2.3, key in attaining optimal regret is using the right level $r \in \mathcal{R}$. An immediate difficulty is that the oracle choice of level used in Definition 7 depends on the unknown significant phase length $\tau_{i+1} - \tau_i$. To circumvent this, as in previous works [Perchet and Rigollet, 2013, Slivkins, 2014], we rely on an adaptive time-varying choice of level r_t. Specifically, each base algorithm choose the level $r_{t-t_{\text{start}}}$ based on the time elapsed since the time t_{start} it was first activated.

Sharing Information across Base Algorithms. Instances of Base-Alg and CMETA share information, in the form of *global variables* as listed below:

- All variables defined in CMETA, namely $t_\ell, t, \{\mathcal{A}_{\text{master}}(B)\}_{B\in\mathcal{T}}, \{Z_{m,t}\}$ (see Lines 3–6 of Algorithm 1).
- All arms played at any round t, along with observed rewards Y_t^a, and the candidate arm set $\mathcal{A}_t$ which takes the value of the set $\mathcal{A}(B)$ of the active Base-Alg at round t and bin $B = T_r(X_t)$ used.

By sharing these global variables, any Base-Alg can trigger a new episode: every time an arm is evicted from $\mathcal{A}(B)$ a Base-Alg , it is also evicted from $\mathcal{A}_{\text{master}}(B)$, which is essentially the candidate arm set for the current episode. A new episode is triggered at time t when $\mathcal{A}_{\text{master}}(B)$ becomes empty for some bin B (necessarily a currently experienced bin), i.e., there is no *safe* arm left to play at the context X_t in the sense of Definition 6.

Note that $\mathcal{A}(B)$ are *local variables* internal to each Base-Alg (the owner of which will be clear from context in usage).

To ensure consistent behavior while using a time-varying choice of level, we enforce further regularity in arm evictions across $\mathcal{X}$: arms evicted from $\mathcal{A}(B')$ are also evicted from child bins $B \subseteq B'$ to ensure $\mathcal{A}(B) \subseteq \mathcal{A}(B')$.

Estimating Aggregate Local Gaps. The quantity $\sum_{s=s_1}^{s_2} \delta_s(a', a) \cdot \mathbf{1}\{X_s \in B\}$ is estimated as $\sum_{s=s_1}^{s_2} \hat{\delta}_s^B(a', a)$, whereby the relative gap $\delta_s(a', a) \cdot \mathbf{1}\{X_s \in B\}$ is estimated by importance weighting as:

$$\hat{\delta}_s^B(a', a) \doteq |\mathcal{A}_t| \cdot \Big(Y_t^{a'} \cdot \mathbf{1}\{\pi_t = a'\} - Y_t^a \cdot \mathbf{1}\{\pi_t = a\} \Big) \cdot \mathbf{1}\{a \in \mathcal{A}_t\} \cdot \mathbf{1}\{X_s \in B\}. \tag{4}$$

Note that the above is an unbiased estimate of $\delta_t(a', a) \cdot \mathbf{1}\{X_s \in B\}$ whenever a' and a are both in $\mathcal{A}_t$ at time t, conditional on the contexts X_t. It then follows that, conditional on $\mathbf{X}_T$, the difference $\sum_{t=s_1}^{s_2} \Big(\hat{\delta}_t^B(a', a) \cdot \mathbf{1}\{X_s \in B\} - \delta_t(a', a) \Big)$ is a martingale that concentrates at a rate roughly $\sqrt{K \cdot n_B([s_1, s_2])}$, where recall from earlier that $n_B(I) \doteq \sum_{s \in I} \mathbf{1}\{X_s \in I\}$ is the context count in bin B over interval I.

ut An arm a is then evicted at round t if, for some fixed $C_0 > 0$ [1], $\exists$ rounds $s_1 < s_2 \leq t$ such that at level $r_{s_2-s_1}$ and (i.e., the bin at level $r_{s_2-s_1}$ containing X_t) letting $B := T_{s_2-s_1}(X_t)$ (i.e., the bin at level $r_{s_2-s_1}$ containing X_t)

$$\max_{a' \in [K]} \sum_{s=s_1}^{s_2} \hat{\delta}_s^B(a', a) > \log(T) \sqrt{C_0 \cdot (K n_B([s_1, s_2]) \vee K^2)} + r_{s_2-s_1} \cdot n_B([s_1, s_2]). \tag{5}$$

[1] $C_0 > 0$ needs to be sufficiently large, but is a universal constant free of the horizon T or any distributional parameters.

5 Key Technical Highlights of Analysis

While a full analysis is deferred to Appendix D due to space constraints, we highlight some of the key novelties and core points of the analysis.

• **Local Safety in Bins implies Safe Total Regret.** We first argue that the notion of significant regret ($\star$) within a bin B captures the total regret rates $T^{\frac{1+d}{2+d}}$ we wish to compete with. If ($\star$) holds for no intervals $[s_1, s_2]$ in all bins B, arm a would be safe and incur little regret over any $[s_1, s_2]$. As it turns out, bounding the per-bin regret by ($\star$) implies a total regret of $T^{\frac{1+d}{2+d}}$ as seen from the following rough calculation: via concentration and the strong density assumption (Assumption 2) to conflate $n_B([1,T]) \approx r(B)^d \cdot T$ and the fact that there are $\approx r^{-d}$ bins at level r, we have:

$$\sum_{B \in T_r} \sqrt{K \cdot n_B([1,T])} + r \cdot n_B([1,T]) \leq K^{1/2} \cdot T^{1/2} \cdot r^{-d/2} + T \cdot r. \tag{6}$$

In particular taking $r \propto (K/T)^{\frac{1}{2+d}}$ makes the above RHS the desired rate $K^{\frac{1}{2+d}} T^{\frac{1+d}{2+d}}$.

• **Significant Regret Threshold is Estimation Error.** At the same time, the RHS of the definition of significant regret ($\star$) is a variance and bias decomposition of the bound on the (conditional on $\mathbf{X}_T$) error of estimating the cumulative regret $\sum_{s=s_1}^{s_2} \delta_s^a(x) \cdot \mathbf{1}\{X_s \in B\}$ at any context $x \in B$. Thus, intuitively, changes of magnitude above the threshold $\sqrt{K \cdot n_B(I)} + r(B) \cdot n_B(I)$ in ($\star$) are detectable.

So, the notion of significant regret ($\star$) perfectly balances both (1) detection of unsafe arms and (2) regret minimization of playing safe arms.

• **A New Balanced Replay Scheduling.** As mentioned earlier in Subsection 3.1, previous adaptive works on contextual bandits fail to attain the optimal regret in this setting due to an inappropriate frequency of scheduling replays. We introduce a novel scheduling (Line 6 of Algorithm 1) which carefully balances exploration and fast detection of significant regret in the sense of ($\star$). The chosen rate $(1/m)^{\frac{1}{2+d}}(1/t)^{\frac{1+d}{2+d}}$ comes from the following intuitive calculation. A scheduled replay of duration m will incur an additional regret of about $m^{\frac{1+d}{2+d}}$. Then, summing over all possible replays, the extra regret incurred due to replays is in total roughly upper bounded by

$$\sum_{t=1}^{T} \sum_{m=2,4,\ldots,T} \left(\frac{1}{m}\right)^{\frac{1}{2+d}} \left(\frac{1}{t}\right)^{\frac{1+d}{2+d}} \cdot m^{\frac{1+d}{2+d}} \lesssim \sum_{t=1}^{T} T^{\frac{d}{2+d}} \cdot (1/t)^{\frac{1+d}{2+d}} \lesssim T^{\frac{1+d}{2+d}}.$$

In other words, the cost of replays only incurs extra constants in the regret. Surprisingly, this scheduling rate is also sufficient for detecting significant regret in *any* experienced subregion B of the context space $\mathcal{X}$, i.e. there is no need to do additional exploration on a localized per-bin basis.

Next, a key feature of the analysis is that one need only minimize regret and detect changes at the critical level $r_{s_2-s_1} \propto (K/(s_2-s_1))^{\frac{1}{2+d}}$. In particular, the following two observations play a major role in bounding the regret.

• **Suffices to Only Check ($\star$) at Critical Levels $r_{s_2-s_1}$.** At first glance, detecting experienced significant shifts (Definition 6) appears difficult as an arm a may incur significant regret over a different bin B' from the bin B that is currently being used by the algorithm.

This difficulty is further compounded by the fact there may even be missing data problems as arms $a \in \mathcal{A}(B)$ in contention at B may have been evicted from sibling bins of the parent $B' \supset B$, thus preventing reliable estimation of a across B'. We in fact show that we only require detecting significant regret in bins B' at the critical level $r_{s_2-s_1}$ and only for the arms still in contention across all of B'. In other words, changes at other levels are all accounted for by changes at this critical level.

Additionally, we observe that the calculations in (6) would hold if we were just concerned with checking ($\star$) for intervals $[s_1, s_2]$ and bins $B_{s_2-s_1}$ at level $r_{s_2-s_1} := \left(\frac{K}{s_2-s_1}\right)^{\frac{1}{2+d}}$. Thus, the critical level $r_{s_2-s_1}$ is the key to both regret minimization and experienced significant shift detection

References

Yasin Abbasi-Yadkori, András György, and Nevena Lazic. A new look at dynamic regret for non-stationary stochastic bandits. *arXiv preprint arXiv:2201.06532*, 2022.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646. PMLR, 22–24 Jun 2014.

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

Sakshi Arya and Yuhong Yang. Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards. *Statistics & Probability Letters*, 164:108818, 2020. ISSN 0167-7152.

Jean-Yves Audibert and Alexander B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. *Conference on Learning Theory*, pages 138–158, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(77):1–40, 2022.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. *AISTATS*, 2011.

Moise Blanchard, Steve Hanneke, and Patrick Jaillet. Non-stationary contextual bandits and universal learning. *arXiv preprint arXiv:2302.07186*, 2023.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *arxiv preprint arXiv:2211.12612*, 2022.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Liyu Chen and Haipeng Luo. Near-optimal goal-oriented reinforcement learning in non-stationary environments. *Advances in Neural Information Processing Systems*, 2022.

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

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pages 1843–1854. PMLR, 2020.

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

Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for CMDPs with non-stationary objectives and constraints. *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 3538–3546. PMLR, 2021.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, page 169–178. AUAI Press, 2011.

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. *Advances in Neural Information Processing Systems*, 33:6743–6754, 2020.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *Proceedings of the 37th International Conference on Machine Learning*, 119: 3199–3210, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1539–1548. PMLR, 10–15 Jul 2018.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for Markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

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

Melody Y Guan and Heinrich Jiang. Nonparametric stochastic contextual bandits. *AAAI*, 2018.

Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. Smoothness-adaptive contextual bandits. *Operations Research*, 70(6):3198–3216, 2022.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. *Conference on Learning Theory*, 2020.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Zohar S Karnin and Oren Anava. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems*, pages 199–207, 2016.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2025–2027. PMLR, 25–28 Jun 2019.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

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

Tyler Lu, Dávid Pál, and Martin Pál. Showing relevant ads via context multi-armed bandits. In *Proceedings of AISTATS*, 2009.

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

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International Conference on Machine Learning*, pages 7447–7458. PMLR, 2021.

Subhojyoti Mukherjee and Odalric-Ambrym Maillard. Distribution-dependent and time-uniform bounds for piecewise i.i.d bandits. *Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Mearning Learning*, 2019.

Ronald Ortner, Pratik Gajane, and Peter Auer. Variational regret bounds for reinforcement learning. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 81–90. PMLR, 22–25 Jul 2020.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022.

Wei Qian and Yuhong Yang. Kernel estimation and model combination in a bandit problem with covariates. *Journal of Machine Learning Research*, 17(149):1–37, 2016a.

Wei Qian and Yuhong Yang. Randomized allocation with arm elimination in a bandit problem with covariates. *Electronic Journal of Statistics*, 10(1):242 – 270, 2016b.

Henry Reeve, Joe Mellor, and Gavin Brown. The k-nearest neighbour ucb algorithm for multi-armed bandits with covariates. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 725–752. PMLR, 07–09 Apr 2018.

Phillipe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *COLT*, 2010.

Jyotirmoy Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, pages 1978–2002, 1991.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: a faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2021.

Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2160–2182. PMLR, 02–05 Jul 2022.

Joseph Suk and Samory Kpotufe. Self-tuning bandits over unknown covariate-shifts. *International Conference on Algorithmic Learning Theory*, 2021.

Ahmed Touati and Pascal Vincent. Efficient learning in non-stationary linear Markov decision processes. *arXiv preprint arXiv:2010.12870*, 2020.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. *Proceedings of the 32nd International Conference on Learning Theory*, 2021.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

Lai Wei and Vaihbav Srivatsva. On abruptly-changing and slowly-varying multiarmed bandit problems. *Annual American Control Conference (ACC)*, 2018.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 495–504, 2018.

Yuhong Yang, Dan Zhu, et al. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

Huozhi Zhou, Jinglin Chen, Lav R. Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

A Details for Specializing Previous Contextual Bandit Results to Lipschitz Contextual Bandits

A.1 Finite Policy Class Contextual Bandits

In the finite policy class setting[2], one is given access to a known finite class Π of policies $\pi : \mathcal{X} \to [K]$, and in the non-stationary variant, seeks to minimize regret to the time-varying benchmark of best policies $\pi_t^* := \text{argmax}_{\pi \in \Pi} \mathbb{E}_{(X,Y)\in\mathcal{D}_t}[Y(\pi(X))]$. In other words, the "dynamic regret" in this setting is defined by (for chosen policies $\{\hat{\pi}_t\}_t$)

$$\mathbb{E}\left[\sum_{t=1}^{T} \max_{\pi\in\Pi} \mathbb{E}_{(X,Y)\in\mathcal{D}_t}[Y(\pi(X))] - \sum_{t=1}^{T} Y_t(\hat{\pi}_t)\right]. \tag{7}$$

We can in fact recover the nonparametric setting and relate the above to our notion of dynamic regret (Definition 2). To do so, we let Π be the class of policies which uses a level $r \in \mathcal{R}$ and discretizes decision-making across individual bins $B \in \mathcal{T}_r$. Then, we claim there is an *oracle sequence of policies* $\{\pi_{\text{t}}^{\text{oracle}}\}_t$ which attains the minimax regret rate of Theorem 1. So, it remains to bound the regret to the sequence $\{\pi_{\text{t}}^{\text{oracle}}\}_t$ in the sense above.

• **Parametrizing in Terms of Global Number L of Shifts.** Suppose there are $L+1$ stationary phases of length $T/(L+1)$. Then, we first claim there is an oracle sequence of policies $\pi_{\text{t}}^{\text{oracle}}$ which attains reget $(L+1)^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}$.

First, recall from Subsection 2.3 the *oracle choice of level* r_n for a stationary period of n rounds, or the level $r_n \propto (K/n)^{\frac{1}{2+d}}$. Now, define $\{\pi_{\text{t}}^{\text{oracle}}\}_t$ as follows: at each round t, $\pi_{\text{t}}^{\text{oracle}}$ uses the oracle level $r := r_{T/(L+1)} \propto \left(\frac{K(L+1)}{T}\right)^{\frac{1}{2+d}}$ and plays in each bin $B \in \mathcal{T}_r$, the arm maximizing the average reward in that bin $\mathbb{E}[f_t^a(X_t) \mid X_t \in B]$. As this is a biased version of the actual bandit problem $\{f_t^a(X_t)\}_{a\in[K]}$ at context X_t, it will follow that $\pi_{\text{t}}^{\text{oracle}}$ incurs regret of order the bias of estimation in B which is r.

Concretely, suppose X_t falls in bin B at level r, and let $\pi_{\text{t}}^{\text{oracle}}(B)$ be the arm selected at round t by $\pi_{\text{t}}^{\text{oracle}}$ in bin B. Then, mean rewards are Lipschitz, each policy $\pi_{\text{t}}^{\text{oracle}}$ suffers regret:

$$\max_{a\in[K]} f_t^a(X_t) - f_t^{\pi_{\text{t}}^{\text{oracle}}(B)}(X_t) \le \max_{a\in[K]} \mathbb{E}[f_t^a(X_t) - f_t^{\pi_{\text{t}}^{\text{oracle}}(B)}(X_t) \mid X_t \in B] + r = r.$$

Thus, the sequence of policies $\{\pi_{\text{t}}^{\text{oracle}}\}_t$ achieves dynamic regret (in the sense of Definition 2)

$$\mathbb{E}\left[\sum_{t=1}^{T} \max_{a\in[K]} f_t^a(X_t) - f_t^{\pi_{\text{t}}^{\text{oracle}}(X_t)}(X_t)\right] \lesssim (L+1)\cdot\left(\frac{T}{L+1}\right)\cdot\left(\frac{K}{(L+1)T}\right)^{\frac{1}{2+d}} \propto L^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}.$$

Thus, it suffices to minimize dynamic regret in the sense of (7) to this oracle policy π_t^{oracle}. The state-of-the-art guarantee in this setting is that of the ADA-ILTCB algorithm of Chen et al. [2019], which achieves a dynamic regret of $\sqrt{KLT\log(|\Pi|)}$. It then remains to compute $|\Pi|$.

As we need only consider levels in $\mathcal{R}$ of size at least $(K/T)^{\frac{1}{2+d}}$, the size of the policy class Π is $|\Pi| = \sum_{r\in\mathcal{R}} K^{r^{-d}} \propto K^{(T/K)^{\frac{d}{2+d}}} \implies \log(|\Pi|) = \left(\frac{T}{K}\right)^{\frac{d}{2+d}} \log(K)$. Plugging this into $\sqrt{KLT\log(|\Pi|)}$ gives a regret rate of $K^{\frac{1}{2+d}} \cdot L^{1/2} T^{\frac{1+d}{2+d}}$, which has a worse dependence on the global number of shifts L than the minimax optimal rate of $L^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}$ (see Theorem 1).

• **Parametrizing in Terms of Total-Variation V_T.** Fix any positive real number $V \in [T^{-\frac{3+d}{2+d}}, T]$. Then, the lower bound construction of Theorem 1 reveals that there exists an environment with $L+1 = T/\Delta$ stationary phases of length $\Delta \doteq \left\lceil \left(\frac{T}{V}\right)^{\frac{2+d}{3+d}} \right\rceil$ and total-variation of order V.

[2]While there are matters of efficiency and what offline learning guarantees may be assumed in this broader *agnostic* setting, we do not discuss these here, and readers are deferred to Langford and Zhang [2008], Dudik et al. [2011], Agarwal et al. [2014].

Then, the earlier defined oracle sequence of policies $\{\pi_t^{\text{oracle}}\}_t$ attains optimal dynamic regret in terms of V_T:

$$(L+1)^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}} \propto T^{\frac{2+d}{3+d}} \cdot V^{\frac{1+d}{3+d}}.$$

Meanwhile, the state-of-the-art regret guarantee in this parametrization is Theorem 2 of Chen et al. [2019], where ADA-ILTCB's regret bound becomes:

$$(K \cdot \log(|\Pi|) \cdot V)^{1/3} T^{2/3} + \sqrt{K \log(|\Pi|) \cdot T} \propto K^{\frac{2}{3(2+d)}} \cdot V^{\frac{1}{3}} \cdot T^{\frac{2+d}{3+d} + \frac{d}{3(2+d)(3+d)}} + K^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}.$$

We claim this rate is worse than our rate in Corollary 5, in fact in all parameters V, K, T. For $K \geq T$, both rates imply linear regret. Assume $K < T$. Then, note by elementary calculations that for all $d \in \mathbb{N} \cup \{0\}$:

$$\frac{2}{3} + \frac{d}{3(2+d)} = \frac{2+d}{3+d} + \frac{1}{3+d} - \frac{2}{3(2+d)}.$$

Then, it follows that rate of Corollary 5 is smaller using the fact that $K < T$:

$$K^{\frac{2}{3(2+d)}} \cdot V^{1/3} \cdot T^{\frac{2}{3} + \frac{d}{3(2+d)}} \geq K^{\frac{2}{3(2+d)}} \cdot V^{\frac{1}{3+d}} \cdot T^{\frac{2+d}{3+d}} \cdot K^{\frac{1}{3+d} - \frac{2}{3(2+d)}} \geq (KV)^{\frac{1}{3+d}} \cdot T^{\frac{2+d}{3+d}}.$$

A.2 Realizable Contextual Bandits

Lipschitz contextual bandits is also a special case of contextual bandits with *realizability*. In this broader setting, the learner is given a function class Φ which contains the true regression function $\phi_t^* : \mathcal{X} \times [K] \to [0,1]$ describing mean rewards of context-arm pairs at round t. The goal is to compete with the time-varying benchmark of policies $\pi_{\phi_t^*}(x) := \text{argmax}_{a \in [K]} \phi_t^*(x, a)$, using calls to a regression oracle over Φ.

While the natural choice for Φ is the infinite class of all Lipschitz functions from $\mathcal{X} \times [K] \to [0,1]$, the state-of-the-art non-stationary algorithm only provides guarantees for finite Φ [Wei and Luo, 2021, Appendix I.7].

However, it is still possible to recover the Lipschitz contextual bandit setting, by defining Φ similarly to how we defined the finite class of policies Π above. Let Φ be the class of all piecewise constant functions which depends on a level $r \in \mathcal{R}$, and are constant on bins $B \in \mathcal{T}_r$ at level r, taking values which are multiples of $T^{-\frac{1}{2+d}}$ (there are $O(T)$ many such values in $[0,1]$). Note this is quite similar to how we defined the policy-class Π above.

For this specification of Φ, the realizability assumption is false. Rather, this is a mildly misspecified regression class which is allowed by the stationary guarantees of FALCON [Simchi-Levi and Xu, 2021, Section 3.2]. In particular, by smoothness, at each round $t \in [T]$ there is a function $\phi_t^* \in \Phi$ such that

$$\sup_{x \in \mathcal{X}, a \in [K]} |\phi_t(x,a) - f_t^a(x)| \lesssim \left(\frac{1}{T}\right)^{\frac{1}{2+d}}.$$

This introduces an additive term in the regret bound of FALCON of order $T^{\frac{1+d}{2+d}}$ which is of the right order in our setting.

Then, the MASTER black-box algorithm using FALCON Simchi-Levi and Xu [2021] as a base algorithm can obtain dynamic regret upper bounded by [see Wei and Luo, 2021, Theorem 2]:

$$\min\left\{\sqrt{\log(|\Phi|) \cdot L \cdot T}, \log^{1/3}(|\Phi|) \cdot \Delta^{1/3} \cdot T^{2/3} + \sqrt{\log(|\Phi|) \cdot T}\right\}.$$

As Φ is essentially the same size as the policy class Π defined in the previous section, the above regret bound specializes to similar rates as those of ADA-ILTCB derived above.

B Useful Lemmas

Throughout the appendix, $c_1, c_2, \ldots$ will denote universal positive constants not depending on T, K or any of the significant shifts $\{\tau_i(\mathbf{X}_T)\}_i$.

B.1 Concentration of Aggregate Gap over an Interval within a Bin

We first recall a Freedman's inequality, which will help us establish concentration of our gap estimators (Proposition 7).

Lemma 6 (Theorem 1 of Beygelzimer et al. [2011])**.** *Let $X_1, \ldots, X_n \in \mathbb{R}$ be a martingale difference sequence with respect to some filtration $\mathcal{F}_0, \mathcal{F}_1, \ldots$. Assume for all t that $X_t \le R$ a.s.. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$\sum_{i=1}^n X_i \le (e-1)\left(\sqrt{\log(1/\delta)\sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_{t-1}]} + R\log(1/\delta)\right). \tag{8}$$

Recall from Section 4 that for round t,

$$\hat{\delta}_t^{a',a}(B) \doteq |\mathcal{A}_t| \cdot (Y_t(a') \cdot \mathbf{1}\{\pi_t = a'\} - Y_t(a) \cdot \mathbf{1}\{\pi_t = a\}) \cdot \mathbf{1}\{a \in \mathcal{A}_t\} \cdot \mathbf{1}\{X_t \in B\}.$$

We next apply Lemma 6 to our aggregate estimator from Section 4.

Proposition 7. *With probability at least $1 - 1/T^2$ w.r.t. the randomness of $\boldsymbol{Y}_T, \{\pi_t\}_t \mid \boldsymbol{X}_T$, we have for all bins $B \in \mathcal{T}$ and rounds $s_1 < s_2$ and all arms $a \in [K]$ that for large enough $c_1 > 0$:*

$$\left|\sum_{s=s_1}^{s_2} \hat{\delta}_s^{i_t,a}(B) - \sum_{s=s_1}^{s_2} \mathbb{E}[\hat{\delta}_s^{a',a}(B)|\mathcal{F}_{s-1}]\right| \le c_1 \log(T)\left(\sqrt{K \cdot n_B([s_1,s_2])} + K\right), \tag{9}$$

where $\mathcal{F} \doteq \{\mathcal{F}_t\}_{t=1}^T$ is the filtration with $\mathcal{F}_t$ generated by $\{\pi_s, Y_s^{\pi_s}\}_{s=1}^t$.

Proof. The proof is similar to the proof of Proposition 3 in Suk and Kpotufe [2022].

The martingale difference $\hat{\delta}_s^{a',a}(B) - \mathbb{E}[\hat{\delta}_s^{a',a}(B) \mid \mathcal{F}_{s-1}]$ is clearly bounded above by $2K$ for all bins B, rounds s, and all arms a, a'. We also have a cumulative variance bound:

$$\begin{aligned}\sum_{s=s_1}^{s_2} \mathbb{E}[(\hat{\delta}_s^{a',a}(B))^2 \mid \mathcal{F}_{s-1}] &\le \sum_{s=s_1}^{s_2} \mathbf{1}\{X_s \in B\} \cdot |\mathcal{A}_s|^2 \cdot \mathbb{E}[\mathbf{1}\{\pi_s = a \text{ or } a'\}|\mathcal{F}_{s-1}] \\ &\le \sum_{s=s_1}^{s_2} \mathbf{1}\{X_s \in B\} \cdot 2|\mathcal{A}_s| \\ &\le 2K \cdot n_B([s_1,s_2]).\end{aligned}$$

Then, the result follows from (8), and taking union bounds over bins B (at most T levels and at most T bins per level), arms a, a', and rounds s_1, s_2. ☐

Since the error probability of Proposition 7 is negligible with respect to regret, we assume going forward in the analysis that (9) holds for all arms $a, a' \in [K]$ and rounds s_1, s_2. Specifically, let $\mathcal{E}_1$ be the good event over which the bounds of Proposition 7 hold for all all arms and intervals $[s_1, s_2]$.

B.2 Concentration of Covariate Counts

Notation. *To ease notation throughout, we'll henceforth use $\mu(\cdot)$ to refer to the context marginal distribution $\mu_X(\cdot)$.*

Lemma 8. *Let $\{i_t\}_{t=1}^T$ be a random sequence of arms whose distribution depends on $\boldsymbol{X}_T$. With probability at least $1 - 1/T^2$ w.r.t. the randomness of $\boldsymbol{X}_T$, we have for all bins $B \in \mathcal{T}$, all arms $a', a \in [K]$, and rounds $s_1 < s_2$, for some large enough $c_2 > 0$ the following inequalities hold:*

$$|n_B([s_1,s_2]) - (s_2 - s_1 + 1) \cdot \mu(B)| \le c_2\left(\log(T) + \sqrt{\log(T)\mu(B) \cdot (s_2 - s_1 + 1)}\right) \tag{10}$$

$$\left|\sum_{s=s_1}^{s_2} \delta_s(i_s, a) \cdot (\mathbf{1}\{X_s \in B\} - \mu_s(B))\right| \le c_2\left(\log(T) + \sqrt{\log(T)\mu(B) \cdot (s_2 - s_1 + 1)}\right) \tag{11}$$

$$\left|\sum_{s=s_1}^{s_2} \delta_s(a) \cdot (\mathbf{1}\{X_s \in B\} - \mu_s(B))\right| \le c_2\left(\log(T) + \sqrt{\log(T)\mu(B) \cdot (s_2 - s_1 + 1)}\right) \tag{12}$$

Proof. The first inequality (10) follow from Lemma 6 since $\sum_{s=s_1}^{s_2} \mathbf{1}\{X_s \in B\} - \mu(B)$ is a martingale, which has conditional variance at most $(s_2 - s_1 + 1) \cdot \mu(B)$.

The other two inequalities are trickier since $\delta_s(a)$ depends on X_s (so that the summand may not be a martingale difference) while $\delta_s(i_s, a)$ may not even be adapted to the canonical filtration generated by $\mathbf{X}_T$ (i.e., i_t may depend on X_s for $s > t$). Nevertheless, we observe that for any random variable $W_s = W_s(\mathbf{X}_T) \in [-1, 1]$:

$$-(\mathbf{1}\{X_t \in B\} - \mu(B)) \leq W_t \cdot (\mathbf{1}\{X_t \in B\} - \mu(B)) \leq \mathbf{1}\{X_t \in B\} - \mu(B).$$

The upper and lower bounds above are both martingale differences with respect to the canonical filtration of $\mathbf{X}_T$ and thus, summing the above over t we have via Lemma 6:

$$\begin{aligned}\left|\sum_{s=s_1}^{s_2} W_s \cdot (\mathbf{1}\{X_t \in B\} - \mu(B))\right| &\leq \left|\sum_{s=s_1}^{s_2} \mathbf{1}\{X_s \in B\} - \mu(B)\right| \\ &\leq c_2 \left(\log(T) + \sqrt{\log(T)\mu(B) \cdot (s_2 - s_1 + 1)}\right).\end{aligned}$$

Then, taking union bounds over rounds s_1, s_2, bins $B \in \mathcal{T}$, and arms $a \in [K]$ gives the result. □

Notation 2 (good event)**.** *Let $\mathcal{E}_1$ be the good event over which the bounds of Proposition 7 hold for all rounds $s_1, s_2 \in [T]$ and arms $a', a \in [K]$. Thus, on $\mathcal{E}_1$, our estimated gaps in each bin will concentrate.*

Let $\mathcal{E}_2$ be the good event on which bounds of Lemma 8 holds for all bins B, arms $a \in [K]$, rounds $s_1, s_2 \in [T]$. Thus, on $\mathcal{E}_2$, our covariate counts $n_B([s_1, s_2])$ will concentrate and we will be able to relate the empirical quantities $\sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mathbf{1}\{X_s \in B\}$ with their expectations.

Next, we establish a lemma which allow us to relate significant regret ($\star$) and thus our eviction criterion (5) between different bins and levels.

Lemma 9 (Relating Aggregate Gaps Between Levels)**.** *On event $\mathcal{E}_2$, if for rounds $s_1 < s_2$, bin B' at level $r_{s_2 - s_1}$ and arm a, for some $c_3 > 0$:*

$$\sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mathbf{1}\{X_s \in B'\} \leq c_3 \left(\sqrt{K \cdot n_{B'}([s_1, s_2]) \vee K^2} + r(B') \cdot n_{B'}([s_1, s_2])\right),$$

then for any bin $B \subseteq B'$ and some $c_4 > 0$:

$$\sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mathbf{1}\{X_s \in B\} \leq c_4 \left(\log^{1/2}(T) \cdot r(B)^d \cdot K^{\frac{1}{2+d}} \cdot (s_2 - s_1)^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T)\mu(B)(s_2 - s_1 + 1)}\right).$$

The same applies for $\delta_s(a)$ replaced with $\delta_s(a', a)$ with any other fixed arm a'.

Proof. We have using (12) and the strong density assumption (Assumption 2):

$$\begin{aligned}\sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mathbf{1}\{X_s \in B\} &\leq \sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mu(B) + c_2 \left(\log(T) + \sqrt{\log(T)(s_2 - s_1 + 1) \cdot \mu(B)}\right) \\ &\leq \frac{r(B)^d}{r(B')^d} \sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mu(B') + c_2 \left(\log(T) + \sqrt{\log(T)(s_2 - s_1 + 1) \cdot \mu(B)}\right)\end{aligned} \tag{13}$$

Again using (12)

$$\begin{aligned}\sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mu_s(B') &\leq \sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mathbf{1}\{X_s \in B'\} + c_2 \left(\log(T) + \sqrt{\log(T)(s_2 - s_1 + 1) \cdot \mu(B')}\right) \\ &\leq c_5 \left(\sqrt{K \cdot n_{B'}([s_1, s_2]) \vee K^2} + r(B') \cdot n_{B'}([s_1, s_2])\right. \\ &\qquad \left. + \log(T) + \sqrt{\log(T)(s_2 - s_1 + 1) \cdot \mu(B')}\right).\end{aligned}$$

Next, applying (10) to $n_{B'}([s_1, s_2])$ and using the strong density assumption (Assumption 2) to bound the mass $\mu(B')$ above by $C_d \cdot r(B')^d$, the above R.H.S. is further upper bounded by

$$c_6 \left(\log^{1/2}(T) K^{\frac{1+d}{2+d}} \cdot (s_2 - s_1)^{\frac{1}{2+d}} + K \log(T) \right). \tag{14}$$

Finally, plugging (14) into (13) and using the fact that $(r(B')/2)^d \geq (K/(s_2 - s_1))^{\frac{d}{2+d}}$, we have that (13) is of the desired order. The proof of the same inequalities with $\delta_s(a', a)$ is analogous. □

The following lemma relating the bias and variance terms in the notion of significant regret ($\star$) will serve useful many places in the analysis. They all follow from concentration and similar calculations via the strong density assumption (Assumption 2) as done previously.

Lemma 10 (bias-variance bound and strong density)**.** *Let $r = r_{s_2 - s_1}$. Then, for any bin $B \in T_r$:*

$$c_7 (s_2 - s_1)^{\frac{1}{2+d}} \cdot K^{\frac{d/2}{2+d}} \leq \sqrt{(s_2 - s_1 + 1) \cdot \mu(B)} \leq c_8 (s_2 - s_1)^{\frac{1}{2+d}} \cdot K^{\frac{d/2}{2+d}}$$
$$\sqrt{n_B([s_1, s_2])} \leq c_9 (s_2 - s_1)^{\frac{1}{2+d}} \cdot K^{\frac{d/2}{2+d}}$$
$$c_{10} (s_2 - s_1)^{\frac{1}{2+d}} \cdot K^{\frac{1+d}{2+d}} \leq n_B([s_1, s_2]) \cdot r \leq c_{11} (s_2 - s_1)^{\frac{1}{2+d}} \cdot K^{\frac{1+d}{2+d}}$$

B.3 Useful Facts about Levels $r \in \mathcal{R}$ and Blocks $[s_\ell(r), e_\ell(r)]$

The following basic facts about the level selection procedure on Line 2 of Algorithm 2 will be useful as we decompose the analysis into the blocks, or different periods of rounds, where different levels are used. The proofs all follow from Notation 1 and basic calculations.

Fact 1 (relating level to interval length)**.** *The level $r_{s_2 - s_1} = 2^{-m}$ satisfies for $s_2 - s_1 \geq K$:*

$$2^{-(m-1)} > \left(\frac{K}{s_2 - s_1} \right)^{\frac{1}{2+d}} \geq 2^{-m},$$

and hence

$$K \cdot 2^{(m-1)(2+d)} < s_2 - s_1 \leq K \cdot 2^{m(2+d)}.$$

Fact 2 (the first block)**.** *The first block $[s_\ell(1), e_\ell(1)]$ consists of rounds $[t_\ell, t_\ell + K]$.*

Fact 3 (start and end times of a block)**.** *For $r < 1$, the* **start time** *or first round $s_\ell(r)$ of the block corresponding to level r in episode $[t_\ell, t_{\ell+1})$ is $s_\ell(r) = t_\ell + \lceil K \cdot (2r)^{-(2+d)} \rceil$ and the* **anticipated end time** *or last round of the block is $e_\ell(r) = t_\ell + \lceil K \cdot r^{-(2+d)} \rceil - 1$.*

Fact 4 (length of a block)**.** *Each block $[s_\ell(r), e_\ell(r)]$ is at least K rounds long. For the first block $[s_\ell(1), e_\ell(1)]$, this is already clear. Otherwise, suppose $r < 1$ in which case:*

$$e_\ell(r) - s_\ell(r) + 1 = \left\lceil K \cdot r^{-(2+d)} \right\rceil - \left\lceil K \cdot (2r)^{-(2+d)} \right\rceil \geq K \cdot r^{-(2+d)} (1 - 2^{-(2+d)}) - 1 \geq K.$$

We also have the above implies

$$2 \cdot (e_\ell(r) - s_\ell(r)) \geq \frac{K \cdot r^{-(2+d)} \cdot (1 - 2^{-(2+d)})}{2}.$$

Rearranging, this becomes for some constant c_{12} depending only on d:

$$c_{12}^{-1} \cdot r \leq \left(\frac{K}{e_\ell(r) - s_\ell(r)} \right)^{\frac{1}{2+d}} < c_{12} \cdot r.$$

Note we can make c_{12} large enough so that the above also holds for level $r = 1$.

The above implies that the block length $e_\ell(r) - s_\ell(r)$ and the episode length $e_\ell(r) - t_\ell(r)$ up to the end of block $[s_\ell(r), e_\ell(r)]$ can be conflated up to constants

$$c_{13}^{-1} \cdot (e_\ell(r) - s_\ell(r)) \leq e_\ell(r) - t_\ell \leq c_{13} \cdot (e_\ell(r) - s_\ell(r)).$$

C Proof of Oracle Regret Bound (Proposition 2)

Recall that $\mathcal{E}_2$ is the good event on which our covariate counts concentrate by Lemma 8. It suffices to show our desired regret bound for any fixed $\mathbf{X}_T$ on this event.

Fix a phase $[\tau_i, \tau_{i+1})$ and let $r = r_{\tau_{i+1}-\tau_i}$. Fix a bin $B \in \mathcal{T}_r$ and let τ_i^a be the last round $t \in [\tau_i, \tau_{i+1})$ such that $X_t \in B$ and arm a is included in $\mathcal{G}_t$. If a is never excluded from $\mathcal{G}_t$ for all such t, let $\tau_i^a \doteq \tau_{i+1} - 1$. WLOG suppose $\tau_i^1 \leq \tau_i^2 \leq \cdots \leq \tau_i^K$. Then, letting B' be the bin at level $r_{\tau_i^a - \tau_i}$ containing covariate $X_{\tau_i^a}$, we have by ($\star$) that:

$$\sum_{t=\tau_i}^{\tau_i^a} \delta_t(a) \cdot \mathbf{1}\{X_t \in B'\} \leq \sqrt{K \cdot n_{B'}([\tau_i, \tau_i^a])} + r(B') \cdot n_{B'}([\tau_i, \tau_i^a]).$$

From Lemma 9, we conclude

$$\sum_{t=\tau_i}^{\tau_i^a} \frac{\delta_t(a) \cdot \mathbf{1}\{X_t \in B\}}{|\mathcal{G}_t|} \leq \frac{c_4 \left(\log^{1/2}(T) \cdot r^d \cdot K^{\frac{1}{2+d}} \cdot (\tau_{i+1}^a - \tau_i)^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T)(\tau_i^a - \tau_i + 1) \cdot \mu(B)} \right)}{K + 1 - a}, \tag{15}$$

where we use the fact that $|\mathcal{G}_t| \geq K + 1 - a$ for $t \leq \tau_i^a$ such that $X_t \in B$. Summing over arms $a \in [K]$ with $\sum_{a \in [K]} \frac{1}{K+1-a} \leq \log(K)$, we obtain:

$$\sum_{a \in [K]} \sum_{t=\tau_i}^{\tau_i^a} \frac{\delta_t(a) \cdot \mathbf{1}\{X_t \in B\}}{|\mathcal{G}_t|} \leq c_4 \log(K) \left(\log^{1/2}(T) r^d K^{\frac{1}{2+d}} (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T)(\tau_i^a - \tau_i + 1) \cdot \mu(B)} \right) \tag{16}$$

Next, we claim that each significant phase $[\tau_i, \tau_{i+1})$ is at least K rounds long or $K \leq \tau_{i+1} - \tau_i$. This follows from the definition of significant regret ($\star$) since for $[s_1, s_2] \subseteq [\tau_i, \tau_{i+1})$:

$$n_B([s_2, s_2]) \geq \sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mathbf{1}\{X_s \in B\} \geq \sqrt{K \cdot n_B([s_1, s_2])} \implies \tau_{i+1} - \tau_i \geq n_B([s_1, s_2]) \geq K.$$

Then $K \leq \tau_{i+1} - \tau_i$ implies (via Fact 1 about the level $r_{\tau_{i+1}-\tau_i}$)

$$\sum_{B \in \mathcal{T}_r} K \log(T) \leq K \log(T) \cdot r^{-d} \leq c_{14} \log(T) K^{\frac{2}{2+d}} (\tau_{i+1} - \tau_i)^{\frac{d}{2+d}} \leq c_{14} \log(T) K^{\frac{1}{2+d}} \cdot (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}}.$$

Additionally, we have by Lemma 10:

$$\sqrt{(\tau_i^a - \tau_i) \cdot \mu(B)} \leq \sqrt{(\tau_{i+1} - \tau_i) \cdot \mu(B)} \leq c_8 (\tau_{i+1} - \tau_i)^{\frac{1}{2+d}} K^{\frac{d/2}{2+d}} \leq c_8 K^{\frac{1}{2+d}} (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}}.$$

Then, plugging the above into (16) and summing over bins B at level r, we have the regret in episode $[\tau_i, \tau_{i+1})$ is with probability at least $1 - 1/T^2$ w.r.t. the distribution of $\mathbf{X}_T$:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=\tau_i}^{\tau_{i+1}-1} \delta_t(\pi_t) \middle| \mathbf{X}_T \right] &= \mathbb{E}\left[\sum_{B \in \mathcal{T}_r} \sum_{t=\tau_i}^{\tau_{i+1}-1} \sum_{a \in \mathcal{G}_t} \frac{\delta_t(a) \cdot \mathbf{1}\{X_t \in B\}}{|\mathcal{G}_t|} \middle| \mathbf{X}_T \right] \\
&= \mathbb{E}\left[\sum_{B \in \mathcal{T}_r} \sum_{a \in [K]} \sum_{t=\tau_i}^{\tau_i^a} \frac{\delta_t(a) \cdot \mathbf{1}\{X_t \in B\}}{|\mathcal{G}_t|} \middle| \mathbf{X}_T \right] \\
&\leq c_{15} \log(K) \sum_{B \in \mathcal{T}_r} \log^{1/2}(T) r^d (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}} + K \log(T) \\
&\leq c_{16} \log(K) \log(T) \cdot (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}},
\end{aligned}$$

where we use the strong density assumption to bound $\sum_{B \in \mathcal{T}_r} r^d \leq \sum_{B \in \mathcal{T}_r} c_d^{-1} \cdot \mu(B) \leq c_d^{-1}$ in the last inequality. Summing the regret over all phases $[\tau_i, \tau_{i+1})$ gives the desired result. $\square$

D Proof of CMETA Regret Upper Bound (Theorem 3)

Recall from Line 3 of Algorithm 1 that t_ℓ is the first round of the ℓ-th episode. WLOG, there are T total episodes and, by convention, we let $t_\ell \doteq T+1$ if only $\ell-1$ episodes occurred by round T.

We first quickly handle the simple case of $T < K$. In this case, the regret bound of Theorem 3 is vacuous since by the sub-additivity of $x \mapsto x^{\frac{1+d}{2+d}}$:

$$\sum_{i=0}^{\tilde{L}} (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}} \geq (\tau_{\tilde{L}+1} - \tau_0)^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}} \geq T^{\frac{1+d}{2+d}} \cdot T^{\frac{1}{2+d}} = T.$$

Thus, it remains to show Theorem 3 for $T \geq K$.

We first transform the expected regret into a more suitable form.

D.1 Decomposing the Regret

It suffices to bound $\mathbb{E}[R_T(\pi, \mathbf{X}_T) \mid \mathbf{X}_T]$ on the good event $\mathcal{E}_1 \cap \mathcal{E}_2$ where the bounds of Lemmas 8 and 9 hold. Going forward in the rest of the analysis, we will assume said bounds hold wherever convenient.

We first transform the regret into a more convenient form. Let $\mathcal{F} \doteq \{\mathcal{F}_t\}_{t=1}^T$ be the filtration with $\mathcal{F}_t$ generated by $\{\pi_s, Y_s^{\pi_s}\}_{s=1}^t$. Then,

$$\begin{aligned}
\mathbb{E}[R_T(\pi, \mathbf{X}_T) \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \mid \mathbf{X}_T] &= \sum_{t=1}^T \mathbb{E}[\mathbb{E}[\delta_t(\pi_t) \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \mid \mathcal{F}_{t-1}] \mid \mathbf{X}_T] \\
&= \sum_{t=1}^T \mathbb{E}\left[\sum_{a \in \mathcal{A}_t} \frac{\delta_t(\pi_t)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \middle| \mathbf{X}_T \right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \middle| \mathbf{X}_T \right].
\end{aligned}$$

Next, as alluded to in the oracle procedure (Definition 7), until the end of a significant phase $[\tau_i, \tau_{i+1})$, there is a safe arm in each bin B at level $r_{\tau_{i+1}-\tau_i}$ which is experienced.

Definition 8 (local last safe arm in each phase $a_t^\sharp$)**.** *For a round $t \in [\tau_i, \tau_{i+1})$, let B be the bin at level $r_{\tau_{i+1}-\tau_i}$ which contains X_t and let $t_i(B)$ be the last round in $[\tau_i, \tau_{i+1})$ such that $X_{t_i(B)} \in B$. Then, by Definition 6, there is a* **last safe arm** *$a_t^\sharp$ which does not yet incur significant regret in bin B in the following sense: for all $[s_1, s_2] \subseteq [\tau_i, t_i(B)]$ letting $r = r_{s_2 - s_1}$ and $B' \in \mathcal{T}_r$ such that $B' \supseteq B$ we have:*

$$\sum_{s=s_1}^{s_2} \delta_s(a_t^\sharp) \cdot \mathbf{1}\{X_s \in B'\} < \sqrt{K \cdot n_{B'}([s_1, s_2])} + r \cdot n_{B'}([s_1, s_2]).$$

Remark 4. *The last safe arms $\{a_t^\sharp\}_t$ only depend on the distribution of $\mathbf{X}_T$ and* **not** *on the realized rewards $\mathbf{Y}_T$. In particular, conditional on $\mathbf{X}_T$, they are fixed.*

We first decompose the regret at round t as (a) the regret of $a_t^\sharp$ and (b) the regret of arm a to the last safe arm. In other words, it suffices to bound:

$$\mathbb{E}\left[\sum_{t=1}^T \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \middle| \mathbf{X}_T \right] = \sum_{t=1}^T \delta_t(a_t^\sharp) \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} + \mathbb{E}\left[\sum_{t=1}^T \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a_t^\sharp, a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \middle| \mathbf{X}_T \right].$$

Note that the expectation on the first sum disappears since $a_t^\sharp$ is only a function of $\mathbf{X}_T$ and the mean reward functions $\{f_t^a(\cdot)\}_{t,a}$.

D.2 Bounding the Regret of the Last Safe Arm

Bounding $\sum_{t=1}^{T} \delta_t(a_t^\sharp)$ will be similar to the proof of Proposition 2. We essentially show that the oracle procedure could have also just played arm $a_t^\sharp$ every round.

Fix a phase $[\tau_i, \tau_{i+1})$ and let $r = r_{\tau_{i+1}-\tau_i}$. Fix a bin $B \in \mathcal{T}_r$ and let $a_i(B)$ be the last safe arm $a_t^\sharp$ of the last round $t \in [\tau_i, \tau_{i+1})$ such that $X_t \in B$. Then, $a_t^\sharp = a_i(B)$ for every round $t \in [\tau_i, \tau_{i+1})$ such that $X_t \in B$. Then, we have by Definition 6 that for bin $B' \supseteq B$ at level $r_{t-\tau_i}$:

$$\sum_{s=\tau_i}^{t} \delta_s(a_i(B)) \cdot \mathbf{1}\{X_s \in B'\} \leq \sqrt{K \cdot n_{B'}([\tau_i, t])} + r(B') \cdot n_{B'}([\tau_i, t]).$$

Then, by Lemma 9, we have:

$$\sum_{s=\tau_i}^{t} \delta_s(a_i(B)) \cdot \mathbf{1}\{X_s \in B\} \leq c_4 \left(\log^{1/2}(T) \cdot r^d \cdot (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}} + K \log(T) + \sqrt{\log(T)(t - \tau_i + 1) \cdot \mu(B)} \right). \tag{17}$$

Then, summing the above over bins in the same fashion as the proof of Proposition 2 gives:

$$\sum_{t=\tau_i}^{\tau_{i+1}-1} \delta_t(a_t^\sharp) = \sum_{B \in \mathcal{T}_r} \sum_{s=\tau_i}^{\tau_{i+1}-1} \delta_s(a_i(B)) \cdot \mathbf{1}\{X_s \in B\} \leq c_3 \log(T) \cdot (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}}.$$

Finally, summing over phases $[\tau_i, \tau_{i+1})$ we have $\sum_{t=1}^{T} \delta_t(a_t^\sharp)$ is of the right order.

D.3 Relating Episodes to Significant Phases

We next show that w.h.p. a restart occurs (i.e., a new episode begins) only if a significant shift has occurred sometime within the episode. Recall from Definition 6 that $\tau_1, \tau_2, \ldots, \tau_{\tilde{L}}$ are the times of the significant shifts and that $t_1, \ldots, t_T$ are the episode start times.

Lemma 11 (Restart Implies Significant Shift). *On event $\mathcal{E}_1$, for each episode $[t_\ell, t_{\ell+1})$ with $t_{\ell+1} \leq T$ (i.e., an episode which concludes with a restart), there exists a significant shift $\tau_i \in [t_\ell, t_{\ell+1})$.*

Proof. Fix an episode $[t_\ell, t_{\ell+1})$. Then, by Line 11 of Algorithm 1, there is a bin B such that every arm $a \in [K]$ was evicted from B at some round in the episode, i.e. (5) is true for each arm a on some interval $[s_1, s_2] \subseteq [t_\ell, t_{\ell+1})$. It suffices to show that this implies a significnat shift has occurred between rounds t_ℓ and $t_{\ell+1}$.

Suppose (5) first triggers the eviction of arm a at time t in $B' \supseteq B$ over interval $[s_1, s_2]$ where $r(B') = r_{s_2 - s_1}$. By concentration (9) and our eviction criteria (5), we have that there is an arm $a' \neq a$ such that (using the notation of Proposition 7) for large enough $C_0 > 0$ and some $c_{17} > 0$:

$$\sum_{s=s_1}^{s_2} \mathbb{E}\left[\hat{\delta}_s^B(a', a) \mid \mathcal{F}_{s-1}\right] \geq c_{17} \log(T) \left(\sqrt{K \cdot n_{B'}([s_1, s_2]) + K^2} + r(B') \cdot n_{B'}([s_1, s_2]) \right). \tag{18}$$

Next, if arm a is evicted from $\mathcal{A}(B')$ at round t, then we have by the definition of $\hat{\delta}_s^{B'}(a', a)$ (4):

$$\mathbb{E}[\hat{\delta}_s^{B'}(a', a) \mid \mathcal{F}_{s-1}] = \begin{cases} \delta_s(a', a) \cdot \mathbf{1}\{X_s \in B'\} & a, a' \in \mathcal{A}_s \\ -f_s^a(X_s) \cdot \mathbf{1}\{X_s \in B\} & a \in \mathcal{A}_s, a' \notin \mathcal{A}_s \\ 0 & a \notin \mathcal{A}_s \end{cases}.$$

In any case, the above L.H.S. conditional expectation is bounded above by $\delta_s(a) \cdot \mathbf{1}\{X_s \in B'\}$. Thus, (18) implies arm a incurs significant regret ($\star$) in B' on $[s_1, s_2]$:

$$\sum_{s=s_1}^{s_2} \delta_s(a) \cdot \mathbf{1}\{X_s \in B'\} \geq \sqrt{K \cdot n_{B'}([s_2, s_2])} + r(B') \cdot n_{B'}([s_1, s_2]).$$

Then, since every arm a is evicted in bin B by round t, a significant shift must have occurred between rounds t_ℓ and $t_{\ell+1}$. □

D.4 Regret of CMETA to the Last Safe Arm

It remains to bound $\mathbb{E}[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \delta_t(a_t^\sharp, a)/|\mathcal{A}_t| \mid \mathbf{X}_t]$. We further decompose this sum over t into episodes and then *blocks* where a particular choice of level is used within the episode. The following notation will be useful.

Definition 9. *Let $s_\ell(r)$ and $e_\ell(r)$ denote the first and last rounds when level r is used by the master* Base-Alg *in episode $[t_\ell, t_{\ell+1})$, i.e. rounds $t \in [t_\ell, t_{\ell+1})$ such that $r_{t-t_\ell} = r$. We call $[s_\ell(r), e_\ell(r)]$ a* **block**. *Let $\textsc{Phases}(\ell, r) \doteq \{i \in [\tilde{L}] : [\tau_i, \tau_{i+1}) \cap [s_\ell(r), e_\ell(r)] \neq \emptyset\}$ be the phases which intersect block $[s_\ell(r), e_\ell(r))$, let $T(i, r, \ell) \doteq |[\tau_i, \tau_{i+1}) \cap [s_\ell(r), e_\ell(r)]|$ be the effective length of the phase as observed in block $[s_\ell(r), e_\ell(r)]$.*

Similarly, define $\textsc{Phases}(\ell) \doteq \{i \in [\tilde{L}] : [\tau_i, \tau_{i+1}) \cap [t_\ell, t_{\ell+1}) \neq \emptyset\}$ be the phases which intersect episode $[t_\ell, t_{\ell+1})$.

It will in fact suffice to show w.h.p. w.r.t. the distribution of $\mathbf{X}_T$, for each episode $[t_\ell, t_{\ell+1})$, each block $[s_\ell(r), e_\ell(r)]$ in $[t_\ell, t_{\ell+1})$, and each bin $B \in \mathcal{T}_r$:

$$\mathbb{E}\left[\sum_{t=s_\ell(r)}^{e_\ell(r)} \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a_t^\sharp, a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{X_t \in B\} \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \middle| \mathbf{X}_T\right]$$
$$\leq c_{18} \log^3(T) \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \left(\log(T) + \sum_{i \in \textsc{Phases}(\ell, r)} r(B)^d \cdot T(i, r, \ell)^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}}\right) \middle| \mathbf{X}_T\right] \tag{19}$$

D.5 Summing the Per-(Bin,Block,Episode) Regret over Bins, Blocks, and Episodes.

Admitting (19), we show that the total dynamic regret over T rounds is of the desired order.

Recall from earlier that there are WLOG T total episodes with the convention that $t_\ell \doteq T + 1$ if only ℓ episodes occur by round T. Then, summing our per-bin regret bound (19) over all the bins at level r gives (using strong density to bound $\sum_{B \in r} r^d \leq \frac{C_d}{c_d}$):

$$\mathbb{E}\left[\sum_{B \in \mathcal{T}_r} \sum_{t=s_\ell(r)}^{e_\ell(r)} \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a_t^\sharp, a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{X_t \in B\} \cdot \mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \middle| \mathbf{X}_T\right]$$
$$\leq c_{18} \log^3(T) \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \left(\sum_{B \in \mathcal{T}_r} \log(T) + \sum_{i \in \textsc{Phases}(\ell, r)} T(i, r, \ell)^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}}\right) \middle| \mathbf{X}_T\right]. \tag{20}$$

Next, summing over the different levels r (of which there are at most $\log(T)$ used in any episode), we obtain by Jensen's inequality:

$$\sum_{r \in \mathcal{R}} \sum_{i \in \textsc{Phases}(\ell, r)} T(i, r, \ell)^{\frac{1+d}{2+d}} = \sum_{i \in \textsc{Phases}(\ell)} \sum_{r \in \mathcal{R} : i \in \textsc{Phases}(\ell, r)} T(i, r, \ell)^{\frac{1+d}{2+d}}$$
$$\leq \sum_{i \in \textsc{Phases}(\ell)} \left(\log(T) \sum_{r \in \mathcal{R} : i \in \textsc{Phases}(\ell, r)} T(i, r, \ell)\right)^{\frac{1+d}{2+d}}.$$

Now, we have

$$\sum_{r \in \mathcal{R} : i \in \textsc{Phases}(\ell, r)} T(i, r, \ell) = \sum_{r \in \mathcal{R} : i \in \textsc{Phases}(\ell, r)} |[\tau_i, \tau_{i+1}) \cap [s_\ell(r), e_\ell(r)]| = \tau_{i+1} - \tau_i + 1.$$

We also have (via Fact 1 about level $r_{t_{\ell+1}-t_\ell}$ which is the smallest level used in episode $[t_\ell, t_{\ell+1})$).

$$\begin{aligned}\sum_{r\in\mathcal{R}}\sum_{B\in\mathcal{T}_r}\log(T) &\le \sum_{r\in\mathcal{R}} r^{-d}\cdot\log(T)\\ &\le c_{19}\log^2(T)\left(\frac{t_{\ell+1}-t_\ell}{K}\right)^{\frac{d}{2+d}}\\ &\le c_{20}\log^2(T)\sum_{i\in\textsc{Phases}(\ell)}(\tau_{i+1}-\tau_i)^{\frac{1+d}{2+d}}\cdot K^{\frac{1}{2+d}}.\end{aligned}$$

Thus, combining the above inequalities with (20), we obtain overall bound:

$$c_{18}\log^4(T)\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1\cap\mathcal{E}_2\}\sum_{i\in\textsc{Phases}(\ell)}(\tau_{i+1}-\tau_i)^{\frac{1+d}{2+d}}\cdot K^{\frac{1}{2+d}}\right].$$

Recall now that $\mathcal{E}_1$ is the good event over which the concentration bounds of Proposition 7 hold. Then, using the fact that, on event $\mathcal{E}_1$, each phase $[\tau_i, \tau_{i+1})$ intersects at most two episodes (Lemma 11), summing the above R.H.S over episodes $\ell \in [T]$ gives us (since at most $\log(T)$ blocks per episode) order

$$2\log^4(T)\sum_{i=1}^{\tilde{L}}(\tau_{i+1}-\tau_i)^{\frac{1+d}{2+d}}\cdot K^{\frac{1}{2+d}}.$$

It then remains to show (19).

D.6 Bounding the Per-Bin Per-Block Regret to the Last Safe Arm

To show (19), we first fix a block $[s_\ell(r), e_\ell(r)]$ and a bin $B \in \mathcal{T}_r$. We then further decompose $\delta_t(a_t^\sharp, a)$ in two parts:

(a) The regret of a to the *last local arm*, denoted by $a_r(B)$, to be evicted from $\mathcal{A}_{\text{master}}(B)$ in block $[s_\ell(r), e_\ell(r)]$ (ties are broken arbitrarily).

(b) The regret of the last local arm $a_r(B)$ to the last safe arm $a_t^\sharp$.

In other words, the L.H.S. of (19) is decomposed as:

$$\underbrace{\mathbb{E}\left[\sum_{t=s_\ell(r)}^{e_\ell(r)}\sum_{a\in\mathcal{A}_t}\frac{\delta_t(a_r(B),a)}{|\mathcal{A}_t|}\cdot\mathbf{1}\{X_t\in B\}\middle|\mathbf{X}_T\right]}_{(a)}+\underbrace{\mathbb{E}\left[\sum_{t=s_\ell(r)}^{e_\ell(r)}\delta_t(a_t^\sharp,a_r(B))\cdot\mathbf{1}\{X_t\in B\}\middle|\mathbf{X}_T\right]}_{(b)}.$$

We will show both (a) and (b) are of order (19).

• Bounding the Regret of Other Arms to the Last Local Arm $a_r(B)$. We start by partitioning the rounds t such that $X_t \in B$ and $a \in \mathcal{A}_t$ in (a) according to before or after they are evicted from $\mathcal{A}_{\text{master}}(B)$. Suppose arm a is evicted from $\mathcal{A}_{\text{master}}(B)$ at round $t_r^a \in [s_\ell(r), e_\ell(r)]$ (formally, we let $t_r^a := e_\ell(r)$ if a is not evicted in block $[s_\ell(r), e_\ell(r)]$). Then, it suffices to bound:

$$\mathbb{E}\left[\sum_{a=1}^{K}\sum_{t=s_\ell(r)}^{t_r^a-1}\frac{\delta_t(a_r(B),a)}{|\mathcal{A}_t|}\cdot\mathbf{1}\{X_t\in B\}+\sum_{a=1}^{K}\sum_{t=t_r^a}^{e_\ell(r)}\frac{\delta_t(a_r(B),a)}{|\mathcal{A}_t|}\cdot\mathbf{1}\{a\in\mathcal{A}_t\}\cdot\mathbf{1}\{X_t\in B\}\middle|\mathbf{X}_T\right]. \tag{21}$$

Suppose WLOG that $t_r^1 \le t_r^2 \le \cdots \le t_r^K$. Then, for each round $t < t_r^a$ all arms $a' \ge a$ are retained in $\mathcal{A}_{\text{master}}(B)$ and thus retained in the candidate arm set $\mathcal{A}_t$ for all rounds t where $X_t \in B$. Importantly, at each round t a level of at least r is used since a child Base-Alg can only use a higher level than the master Base-Alg. Thus, $|\mathcal{A}_t| \ge K + 1 - a$ for all $t \le t_r^a$.

Next, we bound the first double sum in (21), i.e. the regret of playing a to $a_r(B)$ from $s_\ell(r)$ to t_r^a. Applying our concentration bounds (Proposition 7), since arm a is not evicted from $\mathcal{A}(B)$ till round t_r^a, on event $\mathcal{E}_1$ we have for some $c_5 > 0$ and any other arm $a' \in \mathcal{A}(B)$ through round $t_r^a - 1$ (i.e., $a' \in \mathcal{A}_t$ for all $t \in [t_\ell, t_r^a)$ such that $X_t \in B$ since we always use level at least r at such a round t): for bin $B' \supseteq B$ at level $r_{t_r^a - 1 - s_\ell(r)}$: on event $\mathcal{E}_1$ (note that we necessarily always have $\mathcal{A}(B') \supseteq \mathcal{A}(B)$ for $B' \supseteq B$):

$$\sum_{t=s_\ell(r)}^{t_r^a - 1} \mathbb{E}[\hat{\delta}_s^{B'}(a', a) \mid \mathcal{F}_{t-1}] \leq c_5 \log(T) \sqrt{K \cdot n_{B'}([s_\ell(r), t_r^a)) \vee K^2} + r(B') \cdot n_{B'}([s_\ell(r), t_r^a)).$$

Next, since $a, a' \in \mathcal{A}_t$ for each $t \in [s_\ell(r), t_r^a - 1)$ such that $X_t \in B$, we have:

$$\forall t \in [s_\ell(r), t_r^a), X_t \in B : \mathbb{E}[\hat{\delta}_t^B(a', a) \mid \mathcal{F}_{t-1}] = \delta_t(a', a).$$

Thus, we conclude

$$\sum_{t=s_\ell(r)}^{t_r^a - 1} \delta_t(a', a) \cdot \mathbf{1}\{X_t \in B\} \leq c_5 \log(T) \sqrt{K \cdot n_{B'}([s_\ell(r), t_r^a)) \vee K^2} + r(B') \cdot n_{B'}([s_\ell(r), t_r^a)).$$

Thus, by Lemma 9, and since $B' \supseteq B$, we conclude for any such a' on event $\mathcal{E}_1$:

$$\sum_{t=s_\ell(r)}^{t_r^a - 1} \frac{\delta_t(a', a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{X_t \in B\} \leq \frac{c_5 \left(\log^{1/2}(T) r^d \cdot K^{\frac{1}{2+d}} \cdot (t_r^a - s_\ell(r))^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T)(\tau_i^1 - \tau_i + 1) \cdot \mu(B)} \right)}{K + 1 - a}, \tag{22}$$

where we use the fact that $|\mathcal{A}_t| \geq K + 1 - a$ for all $t \in [s_\ell(r), t_r^a)$. Since this last bound holds uniformly for all $a' \in \mathcal{A}(B)$ through round $t_r^a - 1$, it must hold for the last master arm $a_r(B)$.

Then, summing over all arms a, we have on event $\mathcal{E}_1$:

$$\sum_{a=1}^{K} \sum_{t=s_\ell(r)}^{t_r^a - 1} \frac{\delta_t(a_r(B), a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{X_t \in B\} \leq c_5 \log(K) \left(\log^{1/2}(T) \cdot r^d \cdot (e_\ell(r) - s_\ell(r))^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}} + \right.$$
$$\left. K \log(T) + \sqrt{\log(T)(t_r^a - s_\ell(r)) \cdot \mu(B)} \right).$$

Note that by Lemma 10:

$$\sqrt{(t_r^a - s_\ell(r)) \cdot \mu(B)} \leq \sqrt{(e_\ell(r) - s_\ell(r)) \cdot \mu(B)} \leq c_8 K^{\frac{d/2}{2+d}} (e_\ell(r) - s_\ell(r))^{\frac{1}{2+d}} \leq c_8 K^{\frac{1}{2+d}} \cdot r^d \cdot (e_\ell(r) - s_\ell(r))^{\frac{1+d}{2+d}}.$$

Thus, it suffices to consider the RHS above as our bound.

Next, we handle the second double sum in (21). We first observe that if arm a is played in bin B after round t_r^a, then it must due to an active replay. The difficulty here is that replays may interrupt each other and so care must be taken in managing the contribution of $\sum_t \delta_t(a_r(B), a)$ (which may be negative) by different overlapping replays.

Our strategy, identical to that of Section B.1 in Suk and Kpotufe [2022], is to partition the rounds when a is played by a replay after round t_r^a according to which replay is active and not accounted for by another replay. This involves carefully designating a subclass of replays whose durations while playing a in B span all the rounds where a is played in B after t_r^a. Then, we cover the times when a is played by a collection of intervals corresponding to the schedules of this subclass of replays, on each of which we can employ the eviction criterion (5) and concentration like before.

For this purpose, we define the following terminology (which is all w.r.t. a fixed arm a):

Definition 10.

(i) For each scheduled and activated Base-Alg (s, m), *let the round $M(s, m)$ be the minimum of two quantities: (a) the last round in $[s, s + m]$ when arm a is retained in $\mathcal{A}(B)$ by* Base-Alg (s, m) *and all of its children, and (b) the last round that* Base-Alg (s, m) *is active and not permanently interrupted. Call the interval $[s, M(s, m)]$ the* **active interval** *of* Base-Alg (s, m).

(ii) Call a replay Base-Alg (s, m) **proper** *if there is no other scheduled replay* Base-Alg (s', m') *such that $[s, s + m] \subset (s', s' + m')$ where* Base-Alg (s', m') *will become active again after round $s + m$. In other words, a proper replay is not scheduled inside the scheduled range of rounds of another replay. Let* PROPER$(s_\ell(r), e_\ell(r))$ *be the set of proper replays scheduled to start in the block $[s_\ell(r), e_\ell(r)]$.*

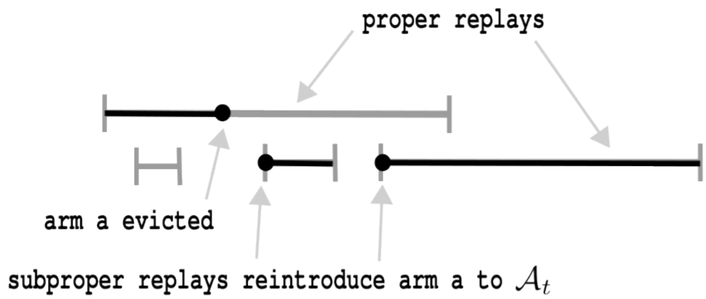

Figure 1: Shown are replay scheduled durations (in gray) with dots marking when arm a is reintroduced to $\mathcal{A}_t$. Black segments indicate the period $[s, M(s,m)]$ for proper and subproper replays. Note that the rounds where $a \in \mathcal{A}_t$ in the left unlabeled replay's duration are accounted for by the larger proper replay.

(iii) Call a scheduled replay $\textsf{Base-Alg}\,(s,m)$ **subproper** *if it is non-proper and if each of its ancestor replays (i.e., previously scheduled replays whose durations have not concluded)* $\textsf{Base-Alg}\,(s',m')$ *satisfies* $M(s',m') < s$. *In other words, a subproper replay either permanently interrupts its parent or does not, but is scheduled after its parent (and all its ancestors) stops playing arm a in B. Let* $\textsc{SubProper}(s_\ell(r), s_\ell(r))$ *be the set of all subproper replays scheduled before round* $t_{\ell+1}$.

Equipped with this language, we now show some basic claims which essentially reduce analyzing the complicated hierarchy of replays to analyzing the active intervals of replays in $\textsc{Proper}(s_\ell(r), e_\ell(r)) \cup \textsc{SubProper}(s_\ell(r), s_\ell(r))$.

Proposition 12. *The active intervals*

$$\{[s, M(s,m)] : \textsf{Base-Alg}\,(s,m) \in \textsc{Proper}(s_\ell(r), e_\ell(r)) \cup \textsc{SubProper}(s_\ell(r), s_\ell(r))\},$$

are mutually disjoint.

Proof. Clearly, the classes of replays $\textsc{Proper}(t_\ell, t_{\ell+1})$ and $\textsc{SubProper}(s_\ell(r), s_\ell(r))$ are disjoint. Next, we show the respective active intervals $[s, M(s,m)]$ and $[s', M(s',m')]$ of any two $\textsf{Base-Alg}\,(s,m)$ and $\textsf{Base-Alg}\,(s',m') \in \textsc{Proper}(s_\ell(r), e_\ell(r)) \cup \textsc{SubProper}(s_\ell(r), s_\ell(r))$ are disjoint.

1. Proper replay vs. subproper replay: a subproper replay can only be scheduled after the round $M(s,m)$ of the most recent proper replay $\textsf{Base-Alg}\,(s,m)$ (which is necessarily an ancestor). Thus, the active intervals of proper replays and subproper replays.
2. Two distinct proper replays: two such replays can only permanently interrupt each other, and since $M(s,m)$ always occurs before the permanent interruption of $\textsf{Base-Alg}\,(s,m)$, we have the active intervals of two such replays are disjoint.
3. Two distinct subproper replays: consider two non-proper replays $\textsf{Base-Alg}\,(s,m), \textsf{Base-Alg}\,(s',m') \in \textsc{SubProper}(s_\ell(r), s_\ell(r))$ with $s' > s$. The only way their active intervals intersect is if $\textsf{Base-Alg}\,(s,m)$ is an ancestor of $\textsf{Base-Alg}\,(s',m')$. Then, if $\textsf{Base-Alg}\,(s',m')$ is subproper, we must have $s' > M(s,m)$, which means that $[s', M(s',m')]$ and $[s, M(s,m)]$ are disjoint.

□

Next, we claim that the active intervals $[s, M(s,m)]$ for $\textsf{Base-Alg}\,(s,m) \in \textsc{Proper}(t_\ell, t_{\ell+1}) \cup \textsc{SubProper}(s_\ell(r), s_\ell(r))$ contain all the rounds where a is played in B after being evicted from $\mathcal{A}_{\text{master}}(B)$. To show this, we first observe that for each round t when a replay is active, there is a unique proper replay associated to t, namely the proper replay scheduled most recently. Next, note that any round $t > t_r^a$ where $X_t \in B$ and where arm $a \in \mathcal{A}_t$ must belong to the active interval $[s, M(s,m)]$ of the unique proper replay $\textsf{Base-Alg}\,(s,m)$ associated to round t, or else satisfies $t > M(s,m)$ in which case a unique subproper replay $\textsf{Base-Alg}\,(s',m') \in \textsc{SubProper}(s_\ell(r), s_\ell(r))$

was active and not yet permanently interrupted by round t. Thus, it must be the case that $t \in [s', M(s', m')]$.

Overloading notation, we'll let $\mathcal{A}_t(B)$ be the value of $\mathcal{A}(B)$ for the Base-Alg active at round t. Next, note that every round $t \in [s, M(s,m)]$ for a proper or subproper Base-Alg (s,m) is clearly a round where $a \in \mathcal{A}_t(B)$ and no such round is accounted for twice by Proposition 12. Thus,

$$\{t \in (t_r^a, e_\ell(r)] : a \in \mathcal{A}_t(B)\} = \bigsqcup_{\text{Base-Alg}\,(s,m) \in \textsc{Proper}(s_\ell(r), e_\ell(r)) \cup \textsc{SubProper}(s_\ell(r), s_\ell(r))} [s, M(s,m)].$$

Then, we can rewrite the second double sum in (21) as:

$$\sum_{a=1}^{K} \sum_{\text{Base-Alg}\,(s,m) \in \textsc{Proper}(s_\ell(r), e_\ell(r)) \cup \textsc{SubProper}(s_\ell(r), s_\ell(r))} Z_{m,s} \cdot \sum_{t=s \vee t_r^a}^{M(s,m)} \frac{\delta_t(a_r(B), a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{X_t \in B\}.$$

Recall in the above that the Bernoulli $Z_{m,s}$ (see Line 6 of Algorithm 1) decides whether Base-Alg (s,m) is scheduled.

Further bounding the sum over t above by its positive part, we can expand the sum over Base-Alg $(s,m) \in \textsc{Proper}(t_\ell, t_{\ell+1}) \cup \textsc{SubProper}(s_\ell(r), s_\ell(r))$ to be over all Base-Alg (s,m), or obtain:

$$\sum_{a=1}^{K} \sum_{\text{Base-Alg}\,(s,m)} Z_{m,s} \cdot \left(\sum_{t=s \vee t_r^a}^{M(s,m)} \frac{\delta_t(a_r(B), a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{X_t \in B\} \right)_+ ,$$

where the sum is over all replays Base-Alg (s,m), i.e. $s \in \{t_\ell + 1, \ldots, t_{\ell+1} - 1\}$ and $m \in \{2, 4, \ldots, 2^{\lceil \log(T) \rceil}\}$. It then remains to bound the contributed relative regret of each Base-Alg (s,m) in the interval $[s \vee t_r^a, M(s,m)]$, which will follow similarly to the previous steps.

We first have using similar arguments as before (now overloading the notation $M(s,m)$ as $M(s,m,a)$ for clarity), i.e. combining our concentration bound (9) with the eviction criterion (5) and applying Lemma 9:

$$\sum_{t=s \vee t_r^a}^{M(s,m)} \frac{\delta_t(a_r(B), a)}{|\mathcal{A}_t|} \cdot \mathbf{1}\{X_t \in B\} \leq \frac{c_5 \left(\log^{1/2}(T) \cdot r^d \cdot K^{\frac{1}{2+d}} \cdot m^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T)(M(s,m) - s)\mu(B)} \right)}{\min_{t \in [s, M(s,m,a)]} |\mathcal{A}_t|}$$

Thus, it remains to bound

$$\sum_{a=1}^{K} \sum_{\text{Base-Alg}\,(s,m)} Z_{m,s} \cdot \left(\frac{c_5 \left(\log^{1/2}(T) \cdot r^d \cdot K^{\frac{1}{2+d}} \cdot m^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T)(M(s,m) - s) \cdot \mu(B)} \right)}{\min_{t \in [s, M(s,m,a)]} |\mathcal{A}_t|} \right).$$

Swapping the outer two sums and recognizing that $\sum_{a=1}^{K} \frac{1}{\min_{t \in [s, M(s,m,a)]} |\mathcal{A}_t|} \leq \log(K)$ by similar arguments to beforeby summing over arms in the order they are evicted by Base-Alg (s,m), we have that it remains to bound

$$\sum_{\text{Base-Alg}\,(s,m)} Z_{m,s} \cdot c_5 \left(\log^{1/2}(T) \cdot r^d \cdot K^{\frac{1}{2+d}} \cdot \tilde{m}^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T)(m - s)\mu(B)} \right), \tag{23}$$

where $\tilde{m} \doteq m \wedge (e_\ell(r) - s_\ell(r))$ (note we may restrict attention to the part of replays in the current block $[s_\ell(r), e_\ell(r)]$). Let

$$R(m, B) \doteq \left(c_5 \left(\log^{1/2}(T) \cdot r^d \cdot K^{\frac{1}{2+d}} \cdot \tilde{m}^{\frac{1+d}{2+d}} + K \log(T) + \sqrt{\log(T) \cdot \tilde{m} \cdot r^d} \right) \right) \wedge n_B([s, s+m]).$$

Then, in light of the previous calculations, $R(m,B)$ is an upper bound on the within-bin B regret contributed by a replay of total duration m (note we can always coarsely upper bound this regret by $n_B([s, s+m])$.

Then, plugging $R(m,B)$ into (23) gives via tower law (we remove the "conditional on $\mathbf{X}_T$" part for ease of presentation):

$$\mathbb{E}\left[\mathbb{E}\left[\sum_{\text{Base-Alg}\,(s,m)} Z_{m,s} \cdot R(m,B) \middle| s_\ell(r) \right] \right] = \mathbb{E}\left[\sum_{s=s_\ell(r)}^{T} \sum_{m} \mathbb{E}[Z_{m,s} \cdot \mathbf{1}\{s \leq e_\ell(r)\} \mid s_\ell(r)] \cdot R(m,B) \right]$$

Next, we observe that $Z_{m,s}$ and $\mathbf{1}\{s \leq e_\ell(r)\}$ are independent conditional on t_ℓ since $\mathbf{1}\{s \leq e_\ell(r)\}$ only depends on the scheduling and observations of base algorithms scheduled before round s. Thus, recalling that $\mathbb{P}(Z_{m,s} = 1) = \mathbb{E}[Z_{m,s} \mid t_\ell] = \left(\frac{1}{m}\right)^{\frac{1}{2+d}} \cdot \left(\frac{1}{s-t_\ell}\right)^{\frac{1+d}{2+d}}$,

$$\begin{aligned}\mathbb{E}[Z_{m,s} \cdot \mathbf{1}\{s \leq e_\ell(r)\} \mid t_\ell] &= \mathbb{E}[Z_{m,s} \mid t_\ell] \cdot \mathbb{E}[\mathbf{1}\{s \leq e_\ell(r)\} \mid s_\ell(r)] \\ &= \left(\frac{1}{m}\right)^{\frac{1}{2+d}} \cdot \left(\frac{1}{s-t_\ell}\right)^{\frac{1+d}{2+d}} \cdot \mathbb{E}[\mathbf{1}\{s \leq e_\ell(r)\} \mid s_\ell(r)].\end{aligned}$$

Plugging this into our expectation from before and unconditioning, we obtain:

$$\mathbb{E}\left[\sum_{s=s_\ell(r)}^{e_\ell(r)} \sum_{n=1}^{\lceil \log(T) \rceil} \left(\frac{1}{2^n}\right)^{\frac{1}{2+d}} \left(\frac{1}{s-t_\ell}\right)^{\frac{1+d}{2+d}} \cdot R(2^n, B)\right] \tag{24}$$

We first evaluate the inner sum over n. Note that

$$\sum_{n=1}^{\lceil \log(T) \rceil} \left(\frac{1}{2^n}\right)^{\frac{1}{2+d}} \cdot (2^n \wedge (e_\ell(r) - s_\ell(r)))^{\frac{1+d}{2+d}} \leq \log(T) \cdot (e_\ell(r) - s_\ell(r))^{\frac{d}{2+d}}$$

$$\sum_{n=1}^{\lceil \log(T) \rceil} \left(\frac{1}{2^n}\right)^{\frac{1}{2+d}} \sqrt{2^n \wedge (e_\ell(r) - s_\ell(r))} \leq (e_\ell(r) - s_\ell(r))^{\frac{d/2}{2+d}}$$

$$\sum_{n=1}^{\lceil \log(T) \rceil} \left(\frac{1}{2^n}\right)^{\frac{1}{2+d}} (K \wedge 2^n) \leq \log(T) \cdot K^{\frac{1+d}{2+d}}.$$

Multiplying by $(s - t_\ell)^{-\frac{1+d}{2+d}}$ and taking a further sum over $s \in [s_\ell(r), e_\ell(r)]$ in the above display, (24) becomes

$$(e_\ell(r) - t_\ell)^{\frac{1}{2+d}} \left((e_\ell(r) - s_\ell(r))^{\frac{d}{2+d}} K^{\frac{1}{2+d}} \cdot r^d + (e_\ell(r) - s_\ell(r))^{\frac{d/2}{2+d}} \sqrt{\log(T) \cdot r^d} + K^{\frac{1+d}{2+d}} \log(T) \right).$$

We have the first term inside the paranetheses above inside dominates the second term as long as $K \geq \log(T)$.

Next, note from Fact 4 that $e_\ell(r) - t_\ell \leq c_{13}(e_\ell(r) - s_\ell(r))$ and so the above is at most:

$$r^d \cdot (e_\ell(r) - s_\ell(r))^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}} + \log(T) K^{\frac{1+d}{2+d}} \cdot (e_\ell(r) - s_\ell(r))^{\frac{1}{2+d}}. \tag{25}$$

We next recall from Fact 4 that each block $[s_\ell(r), e_\ell(r)]$ is at least K rounds long. Thus,

$$C_d \cdot r^d \cdot (e_\ell(r) - s_\ell(r))^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}} \geq c_{21} \cdot (e_\ell(r) - s_\ell(r))^{\frac{1}{2+d}} \cdot K^{\frac{1+d}{2+d}}.$$

Thus, the second term of (25) is at most $\log(T)$ times the first term.

Showing (a) is order (19) then follows from upper bounding $e_\ell(r) - s_\ell(r)$ by the combined length of all phases $[\tau_i, \tau_{i+1})$ intersecting block $[s_\ell(r), e_\ell(r)]$, and using the sub-additivity of $x \mapsto x^{\frac{1+d}{2+d}}$.

• **Bounding the Regret of the Last Master Arm $a_r(B)$ to the Last Safe Arm $a_t^\sharp$.** Before we proceed, we first convert $\sum_{t=s_\ell(r)}^{e_\ell(r)} \delta_t(a_t^\sharp, a_r(B)) \cdot \mathbf{1}\{X_t \in B\}$ into a more convenient form in terms of the bin-masses $\mu(B)$. By concentration (11) of Proposition 7, we have

$$\sum_t \delta_t(a_t^\sharp, a_r(B)) \cdot \mathbf{1}\{X_t \in B\} \leq \sum_t \delta_t(a_t^\sharp, a_r(B)) \cdot \mu(B) + c_1 \left(\log(T) + \sqrt{\log(T)(e_\ell(r) - s_\ell(r)) \cdot \mu(B)} \right).$$

By Lemma 10, we have

$$\sqrt{(e_\ell(r) - s_\ell(r)) \cdot \mu(B)} \leq r^d \cdot (e_\ell(r) - s_\ell(r))^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}}.$$

Additionally, $\log(T)$ is of the right order with respect to (20). Thus, the concentration error terms from Proposition 7 above are negligible.

Moving forward, by the strong density assumption and in light of (19), it suffices to show

$$\sum_{t=s_\ell(r)}^{e_\ell(r)} \delta_t(a_t^\sharp, a_r(B)) \lesssim \sum_{i \in \textsc{Phases}(\ell,r)} (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}}.$$

This is the most difficult quantity to bound since arm $a_t^\sharp$ may have been evicted from $\mathcal{A}_{\text{master}}(B)$ before round t and, thus, we rely on our replay scheduling to bound the regret incurred while waiting to detect a large aggregate value of $\delta_t(a_t^\sharp, a_r(B))$.

For each phase $[\tau_i, \tau_{i+1})$ which intersects the remaining rounds $[s_\ell(r), e_\ell(r)]$ (in an abuse of notation, we'll conflate $e_\ell(r)$ with the *anticipated block end time* based on $s_\ell(r)$; that is, the end block time if no episode restart occurs within the block).

Then, our strategy will be to map out in time the *local bad segments* or subintervals of $[\tau_i, \tau_{i+1})$ where arm $a_r(B)$ incurs significant regret to arm $a_t^\sharp$ in bin B, roughly in the sense of ($\star$). The argument will conclude by arguing that a well-timed replay is scheduled to detect some local bad segment in B, before too many elapse.

As mentioned above, the difficulty here is that $a_r(B)$ is a random variable which depends on all the randomness up to time $e_\ell(r)$. However, conditional on just the block start time $s_\ell(r)$, we define the bad segments for a fixed arm a and then argue that if too many bad segments w.r.t. a elapse in the block, arm a will be evicted in bin B. Crucially, this will hold uniformly over all arms a and thus for arm $a = a_r(B)$, which bounds the regret of $a_r(B)$ in block $[s_\ell(r), e_\ell(r)]$.

Notation. *Going forward, we will drop the dependence on the bin B, level r, block $[s_\ell(r), e_\ell(r)]$, and episode $[t_\ell, t_{\ell+1})$ in certain definitions as they are fixed in the remainder of the analysis. We will let $a_i^\sharp(B)$ denote the last safe of bin B in phase $[\tau_i, \tau_{i+1})$ (see Definition 8).*

Definition 11. *Fix an arm a and $s_\ell(r)$, and let $[\tau_i, \tau_{i+1})$ be any phase intersecting $[s_\ell(r), e_\ell(r)]$. Define rounds $s_{i,0}(a), s_{i,1}(a), s_{i,2}(a) \ldots \in [t_\ell \vee \tau_i, \tau_{i+1})$ recursively as follows: let $s_{i,0}(a) \doteq t_\ell \vee \tau_i$ and define $s_{i,j}(a)$ as the smallest round in $(s_{i,j-1}(a), \tau_{i+1} \wedge e_\ell(r))$ such that arm a satisfies for some fixed $c_{21} > 0$:*

$$\sum_{t=s_{i,j-1}(a)}^{s_{i,j}(a)} \delta_t(a_i^\sharp(B), a) \geq c_{21} \log(T) \cdot (s_{i,j}(a) - s_{i,j-1}(a))^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}}. \tag{26}$$

where $B' \supseteq B$ is the bin at level $r_{s_{i,j}(a) - s_{i,j-1}(a)}$, if such a round $s_{i,j}(a)$ exists. Otherwise, we let the $s_{i,j}(a) \doteq \tau_{i+1} - 1$. We refer to the interval $[s_{i,j-1}(a), s_{i,j}(a))$ as a **bad segment**. *We call $[s_{i,j-1}(a), s_{i,j}(a))$ a* **proper bad segment** *if (26) above holds.*

It will in fact suffice to constrain our attention to proper bad segments, since non-proper bad segments $[s_{i,j-1}(a), s_{i,j}(a))$ (where $s_{i,j}(a) = \tau_{i+1} - 1$ and (26) is reversed) will be negligible in the regret analysis since there is at most one non-proper bad segment per phase $[\tau_i, \tau_{i+1})$ (i.e., the regret of such non-proper bad segments is at most (19)). In what follows, we let $B' \supseteq B$ be the bin at level $r_{s_{i,j}(a) - s_{i,j-1}(a)}$ where $[s_{i,j-1}(a), s_{i,j}(a))$ will be some proper bad segment, known from context.

Lemma 13. *Any proper bad segment is at least K rounds long.*

Proof. We have

$$\begin{aligned}
n_{B'}([s_{i,j}(a), s_{i,j+1}(a)]) &\geq \sum_{s=s_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^\sharp(B), a) \cdot \mathbf{1}\{X_t \in B'\} \\
&\geq \sum_{s=s_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^\sharp(B), a) \cdot \mu(B') - c_2 \left(\log(T) + \sqrt{\log(T)(s_{i,j+1}(a) - s_{i,j}(a)) \cdot \mu(B')} \right) \\
&\geq c_{21} \log(T)(s_{i,j+1}(a) - s_{i,j}(a))^{\frac{1}{2+d}} \cdot K^{\frac{1+d}{2+d}} - c_2 \left(\log(T) + \sqrt{\log(T)(s_{i,j+1}(a) - s_{i,j}(a))\mu(B')} \right) \\
&\geq \sqrt{K \cdot n_{B'}([s_{i,j}(a), s_{i,j+1}(a)])},
\end{aligned}$$

where the last inequality follows from Lemma 10 and choosing c_{21} large enough. $\square$

First, we relate our concentration bound (9) to (26), giving us control of the behavior of CMETA on proper bad segments. But, even before this, we establish an elementary lemma.

Lemma 14. *Let $[s_{i,j}(a), s_{i,j+1}(a))$ be a proper bad segment defined w.r.t. arm a. Let $m \in \mathbb{N}$ be such that $r_{s_{i,j+1}(a)-s_{i,j}(a)} = 2^{-m}$. Then, for some $c_{22} = c_{22}(d) > 0$ depending on the dimension d:*

$$\sum_{t=s_{i,j+1}(a)-K2^{(m-2)(2+d)-1}}^{s_{i,j+1}(a)} \delta_t(a_i^\sharp(B), a) \geq c_{22} \log(T) \cdot K^{\frac{1}{2+d}} \left(s_{i,j+1}(a) - s_{i,j}(a)\right)^{\frac{1+d}{2+d}}. \quad (27)$$

Proof. First, we may assume $s_{i,j+1}(a) - s_{i,j}(a) \geq 4 \cdot K$ by choosing c_4 in (26) large enough (this will make $m - 1$ sensible).

First, observe $K2^{(m-1)(2+d)} \leq s_{i,j+1}(a) - s_{i,j}(a) < K2^{m(2+d)}$. Let $\tilde{s} = s_{i,j+1}(a) - K2^{(m-2)(2+d)-1}$. Then, we have by (26) in the construction of the $s_{i,j}(a)$'s (Definition 11) that:

$$\begin{aligned}\sum_{t=\tilde{s}}^{s_{i,j+1}(a)} \delta_t(a_i^\sharp(B), a) &= \sum_{t=s_{i,j}(a)}^{s_{i,j+1}(a)} \delta_t(a_i^\sharp(B), a) - \sum_{t=s_{i,j}(a)}^{\tilde{s}} \delta_t(a_i^\sharp(B), a) \\ &\geq c_{21} \log(T) K^{\frac{1}{2+d}} \left((s_{i,j+1}(a) - s_{i,j}(a))^{\frac{1+d}{2+d}} - (\tilde{s} - s_{i,j}(a))^{\frac{1+d}{2+d}} \right)\end{aligned}$$

Let $m_{i,j}(a) \doteq s_{i,j+1}(a) - s_{i,j}(a)$. Then, we have

$$m_{i,j}(a) \leq K2^{m(2+d)} \implies \tilde{s} - s_{i,j}(a) = m_{i,j}(a) - K2^{(m-2)(2+d)-1} \leq m_{i,j}(a)(1 - 2^{-2(2+d)-1}).$$

Plugging this into our earlier bound the constants become

$$1 - \left(1 - \frac{1}{2^{2(2+d)+1}}\right)^{\frac{1+d}{2+d}} > 0.$$

Note this last term is positive and only depends on d. □

Lemma 15 (Bin-Count Dominates Concentration Error on Bad Segment)**.** *On event $\mathcal{E}_1$, letting $\tilde{s} = s_{i,j+1}(a) - K2^{(m-2)(2+d)-1}$, we have for bin $B' \supseteq B$ at level $r_{s_{i,j+1}(a)-\tilde{s}}$:*

$$n_{B'}([\tilde{s}, s_{i,j+1}(a)]) \geq 2c_1 \left(\log(T) + \sqrt{\log(T)(s_{i,j+1}(a) - \tilde{s})\mu(B')} \right).$$

Proof. Let $W = s_{i,j+1}(a) - \tilde{s}$. We first claim that $W \geq 2^{-2(2+d)} \cdot (s_{i,j+1}(a) - s_{i,j}(a))$. this follows from $s_{i,j+1}(a) - s_{i,j}(a) \leq K \cdot 2^{m(2+d)}$ and

$$s_{i,j+1}(a) - \tilde{s} = K \cdot 2^{(m-2)(2+d)-1} = 2^{-2(2+d)-1} \cdot (K \cdot 2^{m(2+d)}) \geq 2^{-2(2+d)-1} \cdot (s_{i,j+1}(a) - s_{i,j}(a)).$$

This will allow us to conflate W and $s_{i,j+1}(a) - s_{i,j}(a)$ up to constants.

Since $\overline{\delta}_t^B(a_i^\sharp(B), a) \leq 1$, we have that (27) of the previous lemma and concentration (namely, (11) of Proposition 7; note that although $a_i^\sharp(B)$ is a random variable, it is a fixed and unchanging arm within $[\tau_i, \tau_{i+1})$ and hence $[\tilde{s}, s_{i,j}(a)]$) on $n_{B'}([\tilde{s}, s_{i,j+1}(a)])$ gives

$$\begin{aligned} n_{B'}([\tilde{s}, s_{i,j+1}(a)]) &\geq \sum_{t=\tilde{s}}^{s_{i,j+1}(a)} \delta_t(a_i^\sharp(B), a) \cdot \mathbf{1}\{X_t \in B'\} \\ &\geq c_4 \log(T) \cdot K^{\frac{1+d}{2+d}} \left(s_{i,j+1}(a) - s_{i,j}(a)\right)^{\frac{1}{2+d}} \\ &\geq c_4 \left(\log(T) + \sqrt{\log(T) \cdot W \cdot (K/W)^{\frac{d}{2+d}}} \right) \\ &\geq c_1 \left(\log(T) + \sqrt{\log(T)(s_{i,j+1}(a) - \tilde{s})\mu(B')} \right), \end{aligned}$$

where the last inequality follows from the strong density assumption (Assumption 2). □

Now, we define a well-timed or perfect replay which, if scheduled, will be able to detect the badness of arm a in bin B over a proper bad segment $[s_{i,j}(a), s_{i,j+1}(a))$.

Definition 12 (Perfect Replay). *For a fixed proper bad segment $[s_{i,j}(a), s_{i,j+1}(a))$, define a perfect replay as a* $\textsf{Base-Alg}\,(t_{\text{start}}, m)$ *with* $t_{\text{start}} \in [s_{i,j+1}(a) - K2^{(m-2)(2+d)} + 1, s_{i,j+1}(a) - K2^{(m-2)(2+d)-1}]$ *and* $t_{\text{start}} + m \geq s_{i,j+1}(a)$.

The following proposition analyzes the behavior of a perfect replay and shows it will in fact evict arm a from $\mathcal{A}(B)$ within a proper bad segment $[s_{i,j}(a), s_{i,j+1}(a))$.

Proposition 16. *Suppose event $\mathcal{E}_1$ holds. Let $[s_{i,j}(a), s_{i,j+1}(a))$ be a proper bad segment defined with respect to arm a. Let* $\textsf{Base-Alg}\,(t_{\text{start}}, m)$ *be a perfect replay as defined above which becomes active at t_{start} (i.e., $Z_{t_{\text{start}},m} = 1$). Fix an integer $m \geq s_{i,j+1}(a) - s_{i,j}(a)$. Then:*

(i) $\textsf{Base-Alg}\,(t_{\text{start}}, m)$ **will not** *evict arm $a_i^\sharp(B)$ from $\mathcal{A}(B)$ before round $s_{i,j+1}(a) + 1$ while active.*

(ii) If $a \in \mathcal{A}_t$ for all rounds $t \in [\tilde{s}, s_{i,j+1}(a))$ where $X_t \in B$, where $\tilde{s} = s_{i,j+1}(a) - K2^{(m-2)(2+d)-1}$, then arm a will be excluded from $\mathcal{A}(B)$ by round $s_{i,j+1}(a)$.

Proof. Suppose event $\mathcal{E}_1$ (i.e., our concentration bound (9) holds). For (i), if $a_i^\sharp(B)$ is evicted over $[s_1, s_2] \subseteq [s_{i,j}(a), s_{i,j+1}(a)]$ from $\mathcal{A}(B')$ for bin $B' \supseteq B$ at level $r_{s_2 - s_1}$ by Line 11 of Algorithm 2, then $a_i^\sharp(B)$ incurs significant regret in bin B' over $[s_1, s_2]$ (following same reasoning as in Lemma 11). This is a contradiction to the definition of the last safe arm $a_i^\sharp(B)$ (Definition 8). This shows (i).

For (ii), we first observe $\mathbb{E}[\hat{\delta}_t^B(a_i^\sharp(B), a) \mid \mathcal{F}_{t-1}] = \delta_t(a_i^\sharp(B), a)$ for any round $t \in [\tilde{s}, s_{i,j+1}(a)]$ such that $X_t \in B$ if $a_i^\sharp(B), a \in \mathcal{A}_t$. Let $B' \supseteq B$ be the bin at level $r_{s_{i,j+1}(a) - \tilde{s}}$.

Let $W = s_{i,j+1} - \tilde{s}$. Then, by Lemma 14, we have by smoothness that:

$$\sum_{t=\tilde{s}}^{s_{i,j+1}(a)} \delta_t(a_i^\sharp(B), a) \cdot \mathbf{1}\{X_t \in B'\} \geq c_4 \log(T) K^{\frac{1+d}{2+d}} \cdot W^{\frac{1}{2+d}} - n_{B'}([\tilde{s}, s_{i,j+1}(a)]) \cdot r(B')$$

Next, note that

$$\log(T) \sqrt{K \sum_{s=\tilde{s}}^{s_{i,j+1}} \mu_s(B')} + \log(T) \cdot r(B') \sum_{s=\tilde{s}}^{s_{i,j+1}} \mu_s(B'), \tag{28}$$

is bounded above by the same order.

Next, we bound (28) below by an empirical analogue. Applying concentration on $n_{B'}([\tilde{s}, s_{i,j+1}(a)])$ which dominates the Bernstein error by the previous lemma, the above is further lower bounded by

$$\log(T) \left(\sqrt{K \cdot n_{B'}([\tilde{s}, s_{i,j+1}(a)])} + r(B') \cdot n_{B'}([\tilde{s}, s_{i,j+1}(a)]) \right),$$

meaning arm a will be evicted in B' over $[\tilde{s}, s_{i,j+1}(a)]$.

Furthermore, within $\textsf{Base-Alg}\,(t_{\text{start}}, m)$'s play, arms a and $a_i^\sharp(B)$ will **not** be evicted in any child of B' before round $s_{i,j+1}(a)$ because such an eviction can only happen through a child base algorithm of $\textsf{Base-Alg}\,(t_{\text{start}}, m)$ which will necessarily use a level at least r_W. This is because of the way perfect replays are defined. By definition, the t_{start} is 'close enough" to the critical round $s_{i,j+1}(a) - K2^{(m-2)(2+d)-1}$ so that it will not use a different level than the perfect replay which starts exactly at this critical round.

Formally, we have that the maximum level a perfect replay is $s_{i,j+1}(a) - t_{\text{start}} \leq K \cdot 2^{(m-2)(2+d)} - 1$ and so

$$\left(\frac{K}{s_{i,j+1}(a) - t_{\text{start}}} \right)^{\frac{1}{2+d}} \geq \left(\frac{K}{K \cdot 2^{(m-2)(2+d)} - 1} \right)^{\frac{1}{2+d}} \geq 2^{-(m-2)}.$$

On the other hand,

$$\left(\frac{K}{s_{i,j+1}(a) - \tilde{s}} \right)^{\frac{1}{2+d}} = \frac{1}{2^{m-2-\frac{1}{2+d}}} \in [2^{-(m-2)}, 2^{-(m-3)}).$$

Thus, $2^{-(m-2)} = r_{s_{i,j+1}(a)-\tilde{s}}$ is also the level used to detect that arm a is bad in bin B'. $\square$

Next, we show for any arm a (in particular, $a = a_r(B)$), a perfect replay characterized by Definition 12 is scheduled with high probability if too many bad segments w.r.t. a elapse, thus bounding the regret of a to $a_i^\sharp(B)$ over the phases $[\tau_i, \tau_{i+1})$ intersecting block $[s_\ell(r), e_\ell(r)$.

D.7 Bounding the Regret of the Last Master Arm $a_r(B)$ to the Last Safe Arm $a_t^\sharp$

Next, we bound the the regret of a fixed arm a to $a_i^\sharp(B)$ over the bad segments w.r.t. a in B. it should be understood that in what follows, we condition on $s_\ell(r)$. First, fix an arm a and define the *bad round* $s(a) > s_\ell(r)$ as the smallest round which satisfies, for some fixed $c_{23} > 0$:

$$\sum_{(i,j)} (s_{i,j+1}(a) - s_{i,j}(a))^{\frac{1+d}{2+d}} > c_{23} \log(T)(s(a) - t_\ell)^{\frac{1+d}{2+d}} \tag{29}$$

where the above sum is over all pairs of indices $(i, j) \in \mathbb{N} \times \mathbb{N}$ such that $[s_{i,j}(a), s_{i,j+1}(a))$ is a proper bad segment with $s_{i,j+1}(a) < s(a)$. We will show that arm a is evicted within episode ℓ with high probability by the time the bad round $s(a)$ occurs.

For each proper bad segment $[s_{i,j}(a), s_{i,j+1}(a))$, let $\tilde{s}_{i,j}(a) \doteq s_{i,j+1}(a) - K2^{(m-2)(2+d)-1}$ denote the special point of the bad segment and also let $m_{i,j} \doteq 2^n$ where $n \in \mathbb{N}$ satisfies:

$$2^n \geq s_{i,j+1}(a) - s_{i,j}(a) > 2^{n-1}.$$

Next, recall that the Bernoulli $Z_{m,t}$ decides whether Base-Alg (t, m) activates at round t (see Line 6 of Algorithm 1). If for some $t \in [\hat{s}_{i,j}(a), \tilde{s}_{i,j}(a)]$ where $\hat{s}_{i,j}(a) := s_{i,j+1}(a) - K2^{(m-2)(2+d)} + 1$, $Z_{m_{i,j}.t} = 1$, i.e. a perfect replay is scheduled, then a will be evicted from $\mathcal{A}(B)$ by round $s_{i,j+1}(a)$ (Proposition 16). We will show this happens with high probability via concentration on the sum $\sum_{(i,j)} \sum_t Z_{m_{i,j},t}$ where j, i, t run through all $t \in [\hat{s}_{i,j}(a), \tilde{s}_{i,j}(a))$ and all proper bad segments $[s_{i,j}(a), s_{i,j+1}(a))$ with $s_{i,j+1}(a) < s(a)$. Note that these random variables only depend on the fixed arm a, the block start time $s_\ell(r)$, and the randomness of scheduling replays on Line 6. In particular, the $Z_{m_{i,j},t}$ are independent conditional on t_ℓ.

Then, a Chernoff bound over the randomization of CMETA on Line 6 of Algorithm 1 conditional on t_ℓ yields

$$\mathbb{P}\left(\sum_{(i,j)} \sum_t Z_{m_{i,j},t} \leq \frac{\mathbb{E}[\sum_{(i,j)} \sum_t Z_{m_{i,j},t} \mid s_\ell(r)]}{2} \middle| s_\ell(r) \right) \leq \exp\left(-\frac{\mathbb{E}[\sum_{(i,j)} \sum_t Z_{m_{i,j},t} \mid s_\ell(r)]}{8} \right).$$

We claim the error probability on the R.H.S. above is at most $1/T^3$. To this end, we compute:

$$\mathbb{E}\left[\sum_{(i,j)} \sum_t Z_{m_{i,j},t} \middle| s_\ell(r) \right] \geq \sum_{(i,j)} \sum_{t=\hat{s}_{i,j}(a)}^{\tilde{s}_{i,j}(a)} \left(\frac{1}{m_{i,j}} \right)^{\frac{1}{2+d}} \left(\frac{1}{t - t_\ell} \right)^{\frac{1+d}{2+d}} \geq \frac{1}{4} \sum_{(i,j)} m_{i,j}^{\frac{1+d}{2+d}} \left(\frac{1}{s(a) - t_\ell} \right)^{\frac{1+d}{2+d}} \geq \frac{c_7}{4} \log(T),$$

where the last inequality follows from (29). The R.H.S. above is larger than $24 \log(T)$ for c_{23} large enough, showing that the error probability is small. Taking a further union bound over the choice of arm $a \in [K]$ gives us that $\sum_{(i,j)} \sum_t Z_{m_{i,j},t} > 1$ for all choices of arm a (define this as the good event $\mathcal{E}_3(s_\ell(r))$) with probability at least $1 - K/T^3$.

Recall on the event $\mathcal{E}_1$ the concentration bounds of Proposition 7 hold. Then, on $\mathcal{E}_1 \cap \mathcal{E}_3(s_\ell(r))$, we must have $e_\ell(r) \leq s(a_r(B))$ since otherwise $a_r(B)$ would have been evicted in $\mathcal{A}(B)$ by some perfect replay before the end of the block $e_\ell(r)$ by virtue of $\sum_{(i,j)} \sum_t Z_{m_{i,j},t} > 1$ for arm $a_r(B)$. Thus, by the definition of the bad round $s(a_r(B))$ (29), we must have:

$$\sum_{[s_{i,j}(a_r(B)), s_{i,j+1}(a_r(B))) : s_{i,j+1}(a_r(B)) < e_\ell(r)} (s_{i,j+1}(a_r(B)) - s_{i,j}(a_r(B)))^{\frac{1+d}{2+d}} \leq c_{23} \log(T)(e_\ell(r) - t_\ell)^{\frac{1+d}{2+d}} \tag{30}$$

Thus, by (26) in Definition 11, over the proper bad segments $[s_{i,j}(a_r(B)), s_{i,j+1}(a_r(B)))$ which elapse before the end of the block $e_\ell(r)$ in phase $[\tau_i, \tau_{i+1})$: the regret is at most

$$\sum_{t=s_\ell(r)}^{e_\ell(r)} \delta_t(a_t^\sharp, a_r(B)) \leq \sum_{(i,j)} \log(T) \cdot K^{\frac{1}{2+d}} m_{i,j}^{\frac{1+d}{2+d}}$$
$$\leq \log^2(T) \cdot K^{\frac{1}{2+d}} \cdot (e_\ell(r) - t_\ell)^{\frac{1+d}{2+d}}$$

Over each non-proper bad segment $[s_{i,j}(a_r(B)), s_{i,j-1}(a_r(B)))$ and the last segment $[s_{i,j}(a_r(B)), e_\ell(r)]$, the regret of playing arm $a_r(B)$ to $a_i^\sharp$ is at most $\log(T) \cdot r(B)^d \cdot K^{\frac{1}{2+d}} m_{i,j}^{\frac{1+d}{2+d}}$ by a similar series of calcuations and since there is at most one non-proper bad segment per phase $[\tau_i, \tau_{i+1})$ (see (26) in Definition 11).

So, we conclude that on event $\mathcal{E}_1 \cap \mathcal{E}_3(s_\ell(r))$:

$$\sum_{t=s_\ell(r)}^{e_\ell(r)} \delta_t(a_t^\sharp, a_r(B)) \leq 2c_{23} \log^2(T) \sum_{i \in \textsc{Phases}(r,\ell)} K^{\frac{1}{2+d}} \cdot (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}}.$$

Taking expectation (all expectations below are conditional on $\mathbf{X}_T$ and the good event $\mathcal{E}_2$ over which we have concentration of covariate counts), we have by conditioning first on $s_\ell(r)$ and then on event $\mathcal{E}_1 \cap \mathcal{E}_3(s_\ell(r))$:

$$\mathbb{E}\left[\sum_{t=s_\ell(r)}^{e_\ell(r)} \delta_t(a_t^\sharp, a_r(B))\right]$$
$$\leq \mathbb{E}_{s_\ell(r)}\left[\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_3(s_\ell(r))\} \sum_{t=s_\ell(r)}^{e_\ell(r)} \delta_t(a_t^\sharp, a_r(B)) \middle| s_\ell(r)\right]\right] + T \cdot \mathbb{E}_{t_\ell}\left[\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1^c \cup \mathcal{E}_2^c(s_\ell(r))\} \middle| s_\ell(r)\right]\right]$$
$$\leq 2c_{23} \log^2(T) \mathbb{E}_{s_\ell(r)}\left[\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_3(t_\ell)\} \sum_{i \in \textsc{Phases}(\ell,r)} K^{\frac{1}{2+d}} (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} \middle| s_\ell(r)\right]\right] + \frac{K}{T^2}$$
$$\leq 2c_{23} \log^2(T) \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1\} \sum_{i \in \textsc{Phases}(\ell,r)} (\tau_{i+1} - \tau_i)^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}}\right] + \frac{1}{T},$$

where in the last step we bound $\mathbf{1}\{\mathcal{E}_1 \cap \mathcal{E}_3(t_\ell)\} \leq \mathbf{1}\{\mathcal{E}_1\}$ and apply tower law again. Plugging this into our earlier concentration bound on $\sum_{t=s_\ell(r)}^{e_\ell(r)} \delta_t(a_t^\sharp, a_r(B)) \cdot \mathbf{1}\{X_t \in B\}$, we conclude this part.
$\square$

E Proof of Corollary 5

The proof of Corollary 5 will follow in a similar fashion to the proof of Corollary 2 in Suk and Kpotufe [2022], which relates the total-variation rates to significant shifts in the non-stationary MAB setting. A novel difficulty here is that our notion of significant shift $\tau_i(\mathbf{X}_T), \tilde{L}(\mathbf{X}_T)$ (Definition 6) depends on the full context sequence $\mathbf{X}_T$, and so it is not clear how the (random) significant phases $[\tau_i(\mathbf{X}_T), \tau_{i+1}(\mathbf{X}_T))$ relate to the total-variation V_T, which is a deterministic quantity.

Our strategy will be to first convert the regret rate of Theorem 3 into one which depends on a weaker *worst-case notion of significant shift* which does not depend on the observed $\mathbf{X}_T$. Although this notion of shift is weaker, it will be easier to relate to the total-variation quantity V_T.

Let $\delta_t^a(x) := \max_{a' \in [K]} f_t^{a'}(x) - f_t^a(x)$ be the gap in mean rewards at the fixed context $x \in \mathcal{X}$.

Definition 13 (worst-case sig shift). *Let $\tau_0 = 1$. Then, recursively for $i \geq 0$, the $(i+1)$-th* **worst-case significant shift** *is recorded at time $\tilde{\tau}_{i+1}$, which denotes the earliest time $\tilde{\tau} \in (\tilde{\tau}_i, T]$ such that there exists $x \in \mathcal{X}$ such that* ***for every arm*** *$a \in [K]$, there exists round $s \in [\tilde{\tau}_i, \tilde{\tau}]$, such that $\delta_s^a(x) \geq \left(\frac{K}{t - \tilde{\tau}_i}\right)^{\frac{1}{2+d}}$.*

We will refer to intervals $[\tilde{\tau}_i, \tilde{\tau}_{i+1}), i \geq 0$, as **worst-case (significant) phases**. *The unknown number of such phases (by time T) is denoted $\tilde{L}_{\text{pop}} + 1$, whereby $[\tilde{\tau}_{\tilde{L}_{\text{pop}}}, \tilde{\tau}_{\tilde{L}_{\text{pop}}+1})$, for $\tau_{\tilde{L}_{\text{pop}}+1} \doteq T + 1$, denotes the last phase.*

We next claim that

$$\mathbb{E}_{\mathbf{X}_T}\left[\sum_{i=0}^{\tilde{L}(\mathbf{X}_T)} (\tau_{i+1}(\mathbf{X}_T) - \tau_i(\mathbf{X}_T))^{\frac{1+d}{2+d}}\right] \leq c_{24} \sum_{i=0}^{\tilde{L}_{\text{pop}}} (\tilde{\tau}_{i+1} - \tilde{\tau}_i)^{\frac{1+d}{2+d}}.$$

This follows since the empirical significant phases $[\tau_i(\mathbf{X}_T), \tau_{i+1}(\mathbf{X}_T))$ interleave the population analogues $[\tilde{\tau}_i, \tilde{\tau}_{i+1})$ in the following sense: at each significant shift $\tau_{i+1}(\mathbf{X}_T)$, for each arm $a \in [K]$, there is around $s \in [\tau_i(\mathbf{X}_T), \tau_{i+1}(\mathbf{X}_T)]$ such that for $\delta_s(X_{\tau_{i+1}}) > \left(\frac{K}{\tau_{i+1} - \tau_i}\right)^{\frac{1}{2+d}}$. This means there must be a worst-case significant shift $\tilde{\tau}_j$ in the interval $[\tau_i(\mathbf{X}_T), \tau_{i+1}(\mathbf{X}_T)]$ since the criterion of Definition 13 is triggered at $x = X_{\tau_{i+1}}$. Thus, by the sub-additivity of the function $x \mapsto x^{\frac{1+d}{2+d}}$. This also allows us to conclude that each worst-case significant phase $[\tilde{\tau}_i, \tilde{\tau}_{i+1})$ can intersect at most two significant phases $[\tau_i(\mathbf{X}_T), \tau_{i+1}(\mathbf{X}_T))$.

Thus,

$$\begin{aligned}
\sum_{i=0}^{\tilde{L}(\mathbf{X}_T)} (\tau_{i+1}(\mathbf{X}_T) - \tau_i(\mathbf{X}_T))^{\frac{1+d}{2+d}} &\leq \sum_{i=0}^{\tilde{L}(\mathbf{X}_T)} \sum_{j:[\tilde{\tau}_j, \tilde{\tau}_{j+1}) \cap [\tau_i(\mathbf{X}_T), \tau_{i+1}(\mathbf{X}_T)) \neq \emptyset} |[\tilde{\tau}_j, \tilde{\tau}_{j+1}) \cap [\tau_i(\mathbf{X}_T), \tau_{i+1}(\mathbf{X}_T))|^{\frac{1+d}{2+d}} \\
&\leq c_{24} \sum_{j=0}^{\tilde{L}_{\text{pop}}} (\tilde{\tau}_{j+1} - \tilde{\tau}_j)^{\frac{1+d}{2+d}},
\end{aligned}$$

where we use Jensen's inequality for $a^p + b^p \leq 2^{1-p}(a+b)^p$ for $p \in (0,1)$ and $a, b \geq 0$ in the last step to re-combine the subintervals of each worst-case significant phase $[\tilde{\tau}_j, \tilde{\tau}_{j+1})$.

Then, it suffices to show

$$\sum_{j=0}^{\tilde{L}_{\text{pop}}} (\tilde{\tau}_{j+1} - \tilde{\tau}_j)^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}} \lesssim T^{\frac{1+d}{2+d}} \cdot K^{\frac{1}{2+d}} + (V_T \cdot K)^{\frac{1}{3+d}} \cdot T^{\frac{2+d}{3+d}}. \tag{31}$$

We first transform the total variation into a more flexible quantity depending on the reward functions $f_t^a(\cdot)$ and the full sequence $\mathbf{X}_T$.

Lemma 17. *Let $G_t : \mathcal{X} \times [0,1]^K \to [-1,1]$ be any measurable function which takes the mean reward vector $f_t : \mathcal{X} \to [0,1]^K$ at round t as input, and outputs a real number in $[-1,1]$. Then, recalling $\mathcal{D}_t$ is the joint distribution of X_t and Y_t, we have for $t = 2, \ldots, T$:*

$$\|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}} \geq \frac{1}{2}\left(G_t(f_t) - G_t(f_{t-1})\right).$$

Proof. This follows from the variational representation of the total variation distance [Polyanskiy and Wu, 2022, Theorem 7], which says for any measurable function $H : \mathcal{X} \times [0,1]^K \to [-1,1]$,

$$\|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}} \geq \frac{1}{2}\left(\mathbb{E}_{(X_t,Y_t)\sim\mathcal{D}_t}[H(X_t, Y_t)] - \mathbb{E}_{(X_{t-1},Y_{t-1})\sim\mathcal{D}_{t-1}}[H(X_{t-1}, Y_{t-1})]\right). \tag{32}$$

In particular, we can take H to only depend on the mean reward functions. □

Now, fix a worst-case significant phase $[\tilde{\tau}_i, \tilde{\tau}_{i+1})$ such that $\tau_{i+1} < T + 1$. By Definition 13, there exists a context $x_i \in \mathcal{X}$ such that for arm $a_i \in \operatorname{argmax}_{a \in [K]} f^a_{\tilde{\tau}_{i+1}}(x_i)$ we have there exists a round $t_i \in [\tau_i, \tau_{i+1}]$ such that:

$$\delta_{t_i}^{a_i}(x_i) > \left(\frac{K}{\tilde{\tau}_{i+1} - \tilde{\tau}_i}\right)^{\frac{1}{2+d}}.$$

On the other hand, $\delta^{a_i}_{\tilde{\tau}_{i+1}}(x_i) = 0$ by the definition of arm a_i being the best at x_i at round $\tilde{\tau}_{i+1}$. Thus,

$$\left(\frac{K}{\tilde{\tau}_{i+1} - \tilde{\tau}_i}\right)^{\frac{1}{2+d}} < \delta^{a_i}_{t_i}(x_i) - \delta^{a_i}_{\tilde{\tau}_{i+1}}(x_i) = \sum_{t=t_i+1}^{\tau_{i+1}} \delta_t(a_i, x_i) - \delta_{t-1}(a_i, x_i).$$

For each round $t = 2, \ldots, T$, let $G_t(f_t) := \delta_t(a_i, x_i)$, where x_i is the associated context of the unique worst-case significant shift $\tilde{\tau}_{i+1}$ such that $t \in [\tilde{\tau}_i, \tilde{\tau}_{i+1})$ and where a_i is defined as above. Then, G_t only depends on the mean reward function $f_t : \mathcal{X} \to [0,1]^K$ at round t and *not* on the observed contexts $\mathbf{X}_T$. Then, since $G_t(\cdot)$ satisfies the condition of Lemma 17, we must have

$$\sum_{i=1}^{\tilde{L}_{\text{pop}}} \left(\frac{K}{\tilde{\tau}_{i+1} - \tilde{\tau}_i}\right)^{\frac{1}{2+d}} < \sum_{t=2}^{T} G_t(f_t) - G_{t-1}(f_{t-1}) \leq \sum_{t=2}^{T} \|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}}. \tag{33}$$

Now, by Hölder's inequality for $p \in (0,1)$ and $q \in \left(0, \frac{1+d}{2+d}\right)$:

$$\sum_{i=1}^{\tilde{L}_{\text{pop}}} (\tilde{\tau}_{i+1} - \tilde{\tau}_i)^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}} \leq T^{\frac{1+d}{2+d}} K^{\frac{1}{2+d}} + \left(\sum_i K^{\frac{1}{2+d}} (\tilde{\tau}_{i+1} - \tilde{\tau}_i)^{-q/p}\right)^p \left(\sum_i K^{\frac{1}{2+d}} (\tilde{\tau}_{i+1} - \tilde{\tau}_i)^{\left(\frac{1+d}{2+d} + q\right) \cdot \frac{1}{1-p}}\right)^{1-p}.$$

In particular, letting $p = \frac{1}{3+d}$ and $q = \frac{1}{(2+d)(3+d)}$ and plugging in our earlier bound (33) makes the above RHS

$$V_T^{\frac{1}{3+d}} \cdot K^{\frac{1}{3+d}} \cdot T^{\frac{2+d}{3+d}}.$$

□

F Proof of Theorem 1

We first note that it suffices to show (3) for integer $L \in [0, T] \cap \mathbb{N}$ as lower bounds for all other L follow via approximation and modifying the constant $c > 0$ in (3). Thus, going forward, fix $V \in [0, T]$ and $L \in \mathbb{Z} \cap [0, T]$.

At a high level, our construction will repeat $L + 1$ a hard environment for stationary contextual bandits. In particular, within each stationary phase of length $T/(L+1)$ one is forced to pay a regret of $\left(\frac{T}{L+1}\right)^{\frac{1+d}{2+d}}$, summing to a total regret lower bound of $(L+1) \cdot \left(\frac{T}{L+1}\right)^{\frac{1+d}{2+d}} \approx L^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}}$.

To get the rate in terms of V in (3), we will choose $L \propto V^{\frac{2+d}{3+d}} \cdot T^{\frac{1}{3+d}}$ appropriately and argue that the total-variation V_T is less than V, so that our constructed environment indeed lies in the family $\mathcal{P}(V, L, T)$. This is similar to the arguments of the analogous lower bound [Besbes et al., 2019, Theorem 1] for the non-contextual non-stationary bandit problem.

We start by establishing a lower bound for stationary Lipschitz context bandits. The construction is identical to that of Rigollet and Zeevi [2010, Theorem 4.1]. We only highlight a minor novelty in circumventing the reliance of the cited result on a positive "margin parameter" $\alpha > 0$.

Proposition 18. *Suppose there are $K = 2$ arms. Then, there exists a stationary Lipschitz contextual bandit environment $\mathcal{E}(n)$ over n rounds such that for any algorithm π taking as input random variable U, independent of $\mathcal{E}(n)$, we have for some constant $c > 0$:*

$$\mathbb{E}_{\mathcal{E}(n), U}[R(\pi, \boldsymbol{X}_T)] \geq c \cdot n^{\frac{1+d}{2+d}}.$$

Proof. Let the covariates X_t be uniformly distributed on $[0,1]^d$ at each round $t \in [n]$, so that $\mu_X \equiv \text{Unif}\{[0,1]^d\}$. For ease of presentation, let us reparametrize the two arms as $+1$ and -1.

At each round $t \in [n]$, let arm -1 have reward $Y_t^{-1} \sim \text{Ber}(1/2)$ and let arm 1 have reward $Y_t^1 \sim \text{Ber}(f(X_t))$ where $f : \mathcal{X} \to [0,1]$ is some function to be defined. Let

$$M := \left\lceil \left(\frac{n}{8e}\right)^{\frac{1}{2+d}} \right\rceil.$$

We next partition $\mathcal{X} = [0,1]^d$ into a regular grid $\mathcal{Q} = \{q_1, \ldots, q_{M^d}\}$, where q_k denotes the center of bin B_k, $k = 1, \ldots, M^d$. Specifically, for each index $\mathbf{k} = (k_1, \ldots, k_d) \in \{1, \ldots, M\}^d$, we define the bin B_k as:

$$B_k = \left\{ x \in \mathcal{X} : \frac{k_\ell - 1}{M} \leq x_\ell \leq \frac{k_\ell}{M}, \ell = 1, \ldots, d \right\}.$$

Define $C_\phi \doteq 1/4$. Then, let $\phi : \mathbb{R}^d \to \mathbb{R}_+$ be a smooth function defined by:

$$\phi(x) = \begin{cases} 1 - \|x\|_\infty & 0 \leq \|x\|_\infty \leq 1 \\ 0 & \|x\|_\infty > 1 \end{cases}.$$

It's straightforward to verify ϕ is 1-Lipschitz over $\mathbb{R}^d$.

Now, define the integer $m = \lceil \mu \cdot M^d \rceil$ where $\mu \in (0,1)$ is chosen small enough to ensure $m \leq M^d$. Define $\Sigma_m = \{-1, 1\}^m$ and for any $\omega \in \Omega_m$, define the function f_ω on $[0,1]^d$ via

$$f_\omega(x) = 1/2 + \sum_{j=1}^{m} \omega_j \cdot \phi_j(x),$$

where $\phi_j(x) \doteq M^{-1} \cdot C_\phi \cdot \phi(M \cdot (x - q_j)) \cdot \mathbf{1}\{x \in B_j\}$. Then, the optimal arm at context $x \in \mathcal{X}$ in this environment is given by $\pi_f^*(x) \doteq \operatorname{sgn}(f(x) - 1/2)$.

Then, define the family $\mathcal{C}$ of environments induced by f_ω for $\omega \in \Omega_m$. Next, let $\operatorname{Int}(B_k)$ be the ℓ_∞ ball centered at q_k of radius $\frac{1}{2M}$. Then, we have for any $x \in \operatorname{Int}(B_k)$,

$$|f_\omega(x) - 1/2| \geq M^{-1} \cdot C_\phi / 2.$$

Then, the worst-case regret over the family of environments in $\mathcal{C}$ is at least

$$\begin{aligned} &\sup_{f \in \mathcal{C}} \mathbb{E} \sum_{t=1}^{n} |f^{(1)}(X_t) - f^{(2)}(X_t)| \cdot \mathbf{1}\{\pi_t(X_t) \neq \pi^*(X_t)\} \\ &\geq \frac{C_\phi}{2M} \sup_{f \in \mathcal{C}} \mathbb{E} \sum_{t=1}^{n} \sum_{j=1}^{m} \mathbf{1}\{\pi_t(X_t) \neq \pi^*(X_t), X_t \in \operatorname{Int}(B_j)\}. \end{aligned}$$

Lower bounding the remaining supremum on the above RHS display by $\Omega(n)$ follows the same steps as the proof of Theorem 4.1 in Rigollet and Zeevi [2010]. In particular, the algorithm π may depend on additional randomness U, independent of the environment, which is ignorable in the KL calculations by use of chain rule. Plugging in the earlier choice of M this makes the above RHS at least $\Omega(n^{\frac{1+d}{2+d}})$.

$\square$

Given Proposition 18, the $(L+1) \cdot \left(\frac{T}{L+1}\right)^{\frac{1+d}{2+d}}$ lower bound immediately follows by constructing a random environment which consists of $L+1$ independent repetitions of the stationary environment $\mathcal{E}(T/(L+1))$. Any such constructed environment clearly has at most L global shifts. Note that the regret over a given stationary phase of length $\frac{T}{L+1}$ is lower bounded by $\left(\frac{T}{L+1}\right)^{\frac{1+d}{2+d}}$ regardless of the information learned prior to that phase, as such information can be formalized as exogeneous randomness U in Proposition 18 w.r.t. the fixed stationary phase.

Next, we tackle the lower bound $V^{\frac{1}{3+d}} \cdot T^{\frac{2+d}{3+d}}$ in terms of total-variation budget V. First, if $V < \left(\frac{1}{T}\right)^{\frac{3+d}{2+d}}$, then we're already done as

$$\left(T^{\frac{1+d}{2+d}} + T^{\frac{2+d}{3+d}} \cdot V^{\frac{1}{3+d}}\right) \wedge \left((L+1)^{\frac{1}{2+d}} T^{\frac{1+d}{2+d}}\right)$$

is minimized by the first term which is of order $T^{\frac{1+d}{2+d}}$. Thus, using Proposition 18 with a single stationary phase $\mathcal{E}(T)$ gives lower bound of the right order. Such an environment clearly has total-variation $V_T = 0 \leq V$.

Let $\Delta \doteq \left\lceil \left(\frac{T}{V}\right)^{\frac{2+d}{3+d}} \right\rceil \leq \left\lceil T^{\frac{1}{3+d}} \right\rceil$ and consider $L+1 = T/\Delta$ stationary phases of length Δ. Then, by the previous arguments we have the regret is lower bounded by

$$(L+1)^{\frac{1}{2+d}} \cdot T^{\frac{1+d}{2+d}} = \frac{T}{\Delta^{\frac{1}{2+d}}} \geq \frac{T}{2^{\frac{1}{3+d}} (T/V)^{\frac{1}{3+d}}} \propto T^{\frac{2+d}{3+d}} \cdot V^{\frac{1+d}{3+d}}.$$

Additionally, $T^{\frac{2+d}{3+d}} \cdot V^{\frac{1+d}{3+d}}$ dominates $T^{\frac{1+d}{2+d}}$ since $V \geq \left(\frac{1}{T}\right)^{\frac{3+d}{2+d}}$. Thus, the regret lower bound is proven in terms of V.

It remains to show the total-variation V_T is at most V in the above constructed environments so that it lies in the family $\mathcal{P}(V, L, T)$.

Clearly, the instantaneous total-variation $\|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}} = 0$ for all rounds t not being the start of a new stationary phase. On the other hand, for such a round t, we have that since conditioning increases the TV [Polyanskiy and Wu, 2022, Theorem 7.5(c)], the instantaneous TV is at most:

$$\|\mathcal{D}_t - \mathcal{D}_{t-1}\|_{\text{TV}} \leq \mathbb{E}_{x \sim \mu_X} \left[\|\mathcal{D}_t(Y_t | X_t = x) - \mathcal{D}_{t-1}(Y_{t-1} | X_{t-1} = x)\|_{\text{TV}} \right].$$

Since $Y_t^a | X_t = x \sim \text{Ber}(f_t^a(x))$, we have the RHS' inner TV quantity is just the total variation between Bernoulli's or $\max_{a \in [2]} |f_t^a(x) - f_{t-1}^a(x)|$. Carefully analyzing the variations int he constructed Lipechitz reward functions in the proof of Proposition 18 reveals this TV between Bernoulli's is at most $\frac{e^{\frac{1}{2+d}}}{8^{\frac{1+d}{2+d}}} \cdot \left(\frac{L+1}{T}\right)^{\frac{1}{2+d}}$ (note the attached constant is < 1 for all $d \in \mathbb{N} \cup \{0\}$).

Summing over phases, we have

$$V_T \leq (L+1)^{\frac{3+d}{2+d}} \cdot T^{-\frac{1}{2+d}} = T \cdot \left(\frac{1}{\Delta}\right)^{\frac{3+d}{2+d}} = V.$$

$\square$