Adaptive Smooth Non-Stationary Bandits

Joe Suk
Columbia University
joe.suk@columbia.edu

Abstract

In a remarkably simple proof, we show that the dynamic regret rates given in terms of *significant switches* in best arm (Suk and Kpotufe, 2022) recover those of the smooth non-stationary bandits, for all smoothness exponents and Hölder norms. This unifies several disparate threads in the literature.

1 Introduction

In the multi-armed bandit (MAB) problem, an agent sequentially chooses actions, from a set of K *arms*, based on partial and uncertain feedback in the form of (bounded) rewards $Y_t(a)$ for past actions $a \in [K]$ (see Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020, for general surveys). The goal is to maximize the cumulative reward.

We consider the non-stationary variant of the problem, where rewards are obliviously adversarial. In particular, we consider a *smooth* reward model of non-stationary MAB where only mild Hölder class assumptions are made on changes in rewards over time. In fact, this model captures any finite-horizon bandit problem (e.g., via a polynomial interpolation). Additionally, the degree of smoothness (as measured by the Hölder exponent or coefficient of the associated Hölder class) can be considered a more fine-grained measure of non-stationarity in comparison to conventional measures appearing in other works on non-stationary MAB. Indeed, the rates in this model smoothly interpolate between the more parametric $\sqrt{LT}$ rates seen in *switching bandits* (Garivier and Moulines, 2011), with L switches in rewards over horizon T, and the $V^{1/3}T^{2/3}$ rates in terms of the *total variation* measure V quantifying magnitude of total changes in rewards (Besbes et al., 2019).

The smooth model has been previously studied in bits and pieces. Most previous works (Slivkins, 2014; Wei and Srivatsva, 2018; Komiyama et al., 2021; Krishnamurthy and Gopalan, 2021) focused on the case of non-stationary rewards which are Lipschitz in time, which is also called *slowly varying* bandits. Recently, Manegueu et al. (2021) studied the more general Hölder continuous rewards with Hölder exponent $\beta \leq 1$ (i.e., the non-differentiable regime), while Jia et al. (2023) studied differentiable Hölder reward functions.

The known (dynamic[1]) regret upper bounds are scant in these works (see Table 1), even when assuming knowledge of the smoothness. Even more challenging, it remained open whether one could achieve adaptive regret upper bounds without knowing the smoothness. This work resolves these questions and thus unifies these disparate threads in the literature.

Our result is also somewhat surprising since prior approaches (Krishnamurthy and Gopalan, 2021; Manegueu et al., 2021; Jia et al., 2023) more or less relied on confidence intervals on the magnitude of change in rewards, whose design requires knowledge of the smoothness. In non-parametric statistics, it's long been known that it's impossible to design confidence intervals adaptive to unknown smoothness (Low, 1997), ruling out approaches of this kind.

[1]as measured to a time-varying sequence of best arms.

1.1 Further Discussion on Related Works

Smooth Non-Stationary Bandits. To our knowledge, Slivkins (2014) is the first work to study the slowly varying (i.e., Lipschitz rewards in time) bandit problem. Given a bound δ on the drift in rewards between rounds, their Corollary 13 attains $\delta^{1/3} \cdot T$ dynamic regret via a reduction to Lipschitz contextual bandits with deterministic context $X_t \doteq t$. Other works also studied the slowly varying setting getting $\delta^{1/3} \cdot T$ or $\delta^{1/4} \cdot T$ regret (Combes and Proutiere, 2014; Levine et al., 2017; Wei and Srivatsva, 2018; Seznec et al., 2019; Trovò et al., 2020; Komiyama et al., 2021; Ghosh et al., 2022). Some of the mentioned works only used the drift parameter δ as a measure of non-stationarity within more structured bandit problems. Importantly, all of the above works' procedures rely on knowledge of δ. Recently, Krishnamurthy and Gopalan (2021) showed the $\delta^{1/3} \cdot T$ rate is minimax for the class of slowly-varying problems with drift parameter δ.

Manegueu et al. (2021) studied a more general Hölder continuous model where rewards-in-time have Hölder exponent $\beta \in (0, 1]$, and established a regret upper bound with a procedure which requires knowledge of β. Jia et al. (2023) is the first work to study reward functions which are differentiable in time. They derive a dynamic regret lower bound and show matching regret upper bounds for once and twice differentiable reward functions. Once again, all mentioned regret upper bounds crucially rely on knowledge of the smoothness.

Switching and Other Non-Stationary Bandits. Switching bandits was first considered in the adversarial setting by Auer et al. (2002), where a version of EXP3 was shown to attain optimal dynamic regret $\sqrt{LT}$ when tuned with knowledge of the number L of switches. Later works showed similar guarantees in this problem for procedures inspired by stochastic bandit algorithms (Kocsis and Szepesvári, 2006; Yu and Mannor, 2009; Garivier and Moulines, 2011; Mellor and Shapiro, 2013; Liu et al., 2018; Cao et al., 2019). Recently, Auer et al. (2018, 2019); Chen et al. (2019) established the first adaptive and optimal dynamic regret guarantees, without requiring knowledge of L. Other non-stationarity measures, such as the aforementioned total variation, or more nuanced counts than L were recently studied (Suk and Kpotufe, 2022; Abbasi-Yadkori et al., 2023).

Online Learning with Drift. There's also a related thread of works on online learning with drift where the $\delta^{1/3}$ rate appears (Helmbold and Long, 1991; Bartlett, 1992; Helmbold and Long, 1994; Barve and Long, 1997; Long, 1998; Mohri and Muñoz Medina, 2012; Hanneke and Yang, 2019; Mazzetto and Upfal, 2023).

Non-parametric Contextual Bandits. Hölder class assumptions appear broadly in non-parametric statistics (Györfi et al., 2002; Tsybakov, 2009). In particular, Hölder smooth models also naturally appear in the contextual bandit problem (Woodroofe, 1979; Sarkar, 1991; Yang et al., 2002; Lu et al., 2009; Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013; Slivkins, 2014; Qian and Yang, 2016a,b; Reeve et al., 2018; Guan and Jiang, 2018; Gur et al., 2022; Krishnamurthy et al., 2019; Hu et al., 2020; Arya and Yang, 2020; Suk and Kpotufe, 2021; Cai et al., 2024; Suk and Kpotufe, 2023; Blanchard et al., 2023). As mentioned earlier, the smooth non-stationary bandit is in fact a special case of the (stationary) smooth contextual bandit problem when taking the context $X_t \doteq t$.

Interestingly, Gur et al. (2022) show that one cannot in general rate-optimally adapt to unknown smoothness for this problem. As such, adaptive guarantees for this setting are typically made using a self-similarity assumption (Qian and Yang, 2016b; Gur et al., 2022; Cai et al., 2024). However, these results concern random i.i.d. contexts. To contrast, our results for the $X_t \doteq t$ case are fully adaptive to smoothness without requiring self-similarity.

1.2 Contributions

Our contributions are as follows:

(a) We show a dynamic regret lower bound for all Hölder classes of reward functions. New in this work, we give a sharp characterization of the optimal dependence on the number of arms K and Hölder coefficient

λ, which is not considered in the lower bounds of prior works (Krishnamurthy and Gopalan, 2021; Jia et al., 2023).

(b) We next show the META algorithm of Suk and Kpotufe (2022), which attains a dynamic regret bound in terms of so-called *significant switches in best arm*, in fact attains the optimal regret for all Hölder classes without any parameter knowledge.

(c) As a secondary contribution, we study gap-dependent rates for non-stationary bandits. For environments with no significant switch, we propose a new gap-dependent rate based on the idea of a *significant shift oracle* which plays arms until they incur large dynamic regret. We show that this gap-dependent rate recovers a more pessimistic *restarting oracle* gap-dependent rate targeted by prior related works (Mukherjee and Maillard, 2019; Seznec et al., 2020; Krishnamurthy and Gopalan, 2021). We show our new rate is achievable without any parameter knowledge by a randomized elimination algorithm inspired by Suk and Kpotufe (2022). Importantly, this shows that, so long as no significant shift occurs, one can achieve much faster gap-dependent rates than previously thought possible.

(d) Relating this back to the smooth non-stationary bandit, we give a simple and sharp characterization, in terms of the maximum Hölder coefficient, of which smooth bandit models admit these fast gap-dependent regret rates.

Works on Smooth Bandits	Adaptive?	Parameters	Dynamic Regret Upper Bound
Manegueu et al. (2021)	No	$\beta \in (0, 1]$	$T^{\frac{\beta+1}{2\beta+1}} \lambda^{\frac{1}{2\beta+1}} K^{\frac{\beta}{2\beta+1}}$
Slivkins (2014) Krishnamurthy and Gopalan (2021)	No	$\beta = 1$	$T\lambda^{\frac{1}{3}} K^{\frac{1}{3}}$
Jia et al. (2023)	No	$\beta = 1, 2, K = 2$	$T^{\frac{\beta+1}{2\beta+1}} \lambda^{\frac{1}{2\beta+1}}$
This Work (matching upper & lower bounds)	Yes	$\beta > 0$	$T^{\frac{\beta+1}{2\cdot\beta+1}} \lambda^{\frac{1}{2\cdot\beta+1}} K^{\frac{\beta}{2\cdot\beta+1}}$

Table 1: A summary comparison of our dynamic regret bounds with those of prior works.

2 Problem Setup

2.1 Preliminaries and Notation

We assume an oblivious adversary decides a sequence of distributions on the rewards of K arms in $[K]$.

Arm a at round t has random reward $Y_t(a) \in [0, 1]$ with mean $\mu_t(a)$. A (possibly randomized) algorithm π selects at each round t some arm $\pi_t \in [K]$ and observes reward $Y_t(\pi_t)$. The goal is to minimize the *dynamic regret*, i.e., the expected regret to the best arm at each round. This is defined as:

$$R(\pi, T) \doteq \sum_{t=1}^{T} \max_{a\in[K]} \mu_t(a) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_t(\pi_t)\right].$$

We'll use $R_{\mathcal{E}}(\pi, T)$ to denote the expected regret under an environment $\mathcal{E}$.

In this paper, we rely heavily on analyzing the gaps in mean rewards between arms. Thus, let $\delta_t(a', a) \doteq \mu_t(a') - \mu_t(a)$ denote the *relative gap* of arms a to a' at round t. Define the *absolute gap* of arm a as $\delta_t(a) \doteq \max_{a'\in[K]} \delta_t(a', a)$, corresponding to the instantaneous dynamic regret of playing a at round t. Then, the dynamic regret can be written as $\sum_{t\in[T]} \mathbb{E}[\delta_t(\pi_t)]$.

Notation. *Throughout this paper, in theorem statements we'll use $C_0, C_1, \ldots$ to denote universal constants free of $K, T, \beta, \lambda, \{\mu_t(a)\}_{t\in[T],a\in[K]}$. In proofs, universal constants $c_0, c_1, \ldots$ will be used.*

2.2 Smooth Non-Stationary Bandits

We first recall the definition of a Hölder class of functions (Tsybakov, 2009, Definition 1.2).

Definition 1 (Hölder Class Function)**.** *For $\beta, \lambda > 0$, we say a function $f : [0, 1] \to \mathbb{R}$ is (β, λ)-Hölder if f is $m \doteq \lfloor \beta \rfloor$-times differentiable and*

$$\forall x, x' \in [0, 1] : |f^{(m)}(x) - f^{(m)}(x')| \leq \lambda \cdot |x - x'|^{\beta - m}.$$

By convention, we let the zero-th derivative be $f^{(0)}(x) \doteq f(x)$. We call λ the **Hölder coefficient** *whose value may be taken as $\sup_{x \neq x'} \frac{|f^{(m)}(x) - f^{(m)}(x')|}{|x - x'|^{\beta - m}}$.*

Next, we say a bandit environment is Hölder class if the absolute gaps, as functions of normalized time, are (β, λ)-Hölder in the sense above.

Definition 2 (Hölder Gap Environments)**.** *We say a bandit environment is (β, λ)-Hölder if, for every arm $a \in [K]$, there exists a (β, λ)-Hölder function f such that the gap function (in time) is realized by f, i.e. $\delta_t(a) = f(t/T)$ for all $t \in [T]$. We'll use $\Sigma(\beta, \lambda)$ to denote the class of bandit environments which are (β, λ)-Hölder over T rounds.*

We note that, unlike in the aforementioned prior works on smooth non-stationary bandits (Slivkins, 2014; Krishnamurthy and Gopalan, 2021; Manegueu et al., 2021; Jia et al., 2023), our model only relies on characterizing the smoothness of the absolute gap functions $\delta_t(a)$ in time t, and not on the reward functions $\mu_t(a)$. In particular, changes in rewards **can be arbitrarily rough** and changes in rewards $\mu_t(a)$ which do not change the gaps $\delta_t(a)$ do not enter into our regret rates.

3 Dynamic Regret Lower Bound

We first characterize the minimax regret rate over the class of problems in $\Sigma(\beta, \lambda)$. For comparison, Jia et al. (2023, Theorem 3.4) already established a lower bound for integer smoothness $\beta \in \mathbb{Z}_{\geq 1}$, $K = 2$ arms, and fixed Hölder coefficient $\lambda = 1$. Our main novelty here is to show a more comprehensive lower bound which captures sharp dependence on all of T, K, λ.

Theorem 3. *(Proof in Appendix A) Fix $\beta, \lambda > 0$, $K \geq 2$, and $T \in \mathbb{N}$. For any algorithm π, there exists an environment $\mathcal{E} \in \Sigma(\beta, \lambda)$ such that the regret is lower bounded by*

$$R_{\mathcal{E}}(\pi, T) \geq \Omega(\min\{\sqrt{KT} + T^{\frac{\beta+1}{2\cdot\beta+1}} \cdot \lambda^{\frac{1}{2\cdot\beta+1}} \cdot K^{\frac{\beta}{2\cdot\beta+1}}, T\}).$$

Note that if the gap functions $t \mapsto \delta_t(\cdot)$ are C^∞ smooth in time, the above rate of $T^{\frac{\beta+1}{2\beta+1}} \cdot \lambda^{\frac{1}{2\beta+1}}$ for (β, λ)-Hölder gaps becomes $T^{1/2}$ as $\beta \to \infty$. Thus, the rate of Theorem 3 interpolates the stationary regret rate $\sqrt{T}$ and the $T^{2/3}$ regret seen in slowly-varying $\beta = 1$ bandits (Krishnamurthy and Gopalan, 2021).

Remark 1. *For $\beta = 1$ (i.e., the slowly-varying setting), the above rate becomes $T^{2/3}\lambda^{1/3}K^{1/3} = T \cdot (\lambda/T)^{1/3} \cdot K^{1/3}$ which is the rate seen in Slivkins (2014) for drift parameter $\delta = \lambda/T$.*

4 Dynamic Regret Upper Bound

As alluded in Subsection 1.2, our main dynamic regret upper bound is achieved by the META algorithm of Suk and Kpotufe (2022) which adapts to so-called *significant shifts in best arm*. The key idea behind this result is that a significant shift encodes large variation with respect to any (β, λ)-Hölder environment, thus allowing us to recover the rate of Theorem 3. We now recall the notion of a significant shift.

First, we say arm a incurs **significant regret**[2] on interval[3] $[s_1, s_2]$ if:

$$\sum_{t=s_1}^{s_2} \delta_t(a) \geq \sqrt{K \cdot (s_2 - s_1 + 1)}, \tag{1}$$

or intuitively if it incurs large dynamic regret. On the other hand, if (1) holds for no interval in a time period, then arm a incurs little regret over that period and is *safe* to play. Thus, a *significant shift* is recorded only when there is no safe arm left to play. The following recursive definition captures this.

Definition 4. *Let $\tau_0 = 1$. Then, recursively for $i \geq 0$, the $(i+1)$-th* **significant shift** *is recorded at time τ_{i+1}, which denotes the earliest time $\tau \in (\tau_i, T]$ such that* for every arm $a \in [K]$, *there exists rounds $s_1 < s_2, [s_1, s_2] \subseteq [\tau_i, \tau]$, such that arm a has significant regret (1) on $[s_1, s_2]$.*

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

Then, a *significant shift oracle*, which roughly plays arms until they're unsafe in each significant phase and then restarts at each significant shift, attains a regret bound of (Suk and Kpotufe, 2022, Proposition 1)

$$\sum_{i=0}^{\tilde{L}} \sqrt{K \cdot (\tau_{i+1} - \tau_i)}. \tag{2}$$

The main result of Suk and Kpotufe (2022) is to match the above rate up to log terms without any knowledge of non-stationarity. Their META algorithm estimates when the significant shifts τ_i occur using importance-weighted estimates of the gaps and then restarts an elimination procedure upon detecting an empirical version of a significant shift. The inner workings of the algorithm are beyond the scope of this discussion and surprisingly irrelevant for our result here; the interested reader is deferred to Section 4 of Suk and Kpotufe (2022).

Our main result is that, independent of the algorithm and for any environment, the regret rate (2) inherently captures the minimax rate for smooth non-stationary bandits.

Theorem 5 (Proof in Appendix B)**.** *Consider any (β, λ)-Hölder environment over T rounds and let $\{\tau_i\}_{i=0}^{\tilde{L}}$ be the significant shifts of the environment as in Definition 4. Then, we have:*

$$\sum_{i=0}^{\tilde{L}} \sqrt{K \cdot (\tau_{i+1} - \tau_i)} \leq C_0 \sqrt{\beta + 1} \left(\sqrt{KT} + T^{\frac{\beta+1}{2\beta+1}} \cdot \lambda^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}} \right).$$

An immediate corollary is that the META algorithm can match the lower bound of Theorem 3 up to log terms.

Corollary 6. *By Theorem 1 of Suk and Kpotufe (2022), the* META *algorithm has an expected regret upper bound:*

$$R(\pi, T) \leq C_1 \log(K) \log^2(T) \sqrt{\beta + 1} \left(\sqrt{KT} + T^{\frac{\beta+1}{2\beta+1}} \cdot \lambda^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}} \right).$$

Remark 2. *Note that any non-stationary bandit environment over T rounds can be captured by a (β, λ)-Hölder environment for any $\beta > 0$ using, e.g., a Lagrange interpolation of the finite data $\{(t/T, \mu_t(a))\}_{t \in [T], a \in [K]}$. As $T^{\frac{\beta+1}{2\beta+1}} \cdot \lambda^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}} \to \sqrt{KT}$ as $\beta \to \infty$, this seems to suggest we can recover a stationary $\sqrt{KT}$ regret rate for any non-stationary environment, which is seemingly a contradiction. However, taking $\beta \to \infty$ will make the bound of Theorem 5 vacuous as there is a constant dependence of $\sqrt{\beta + 1}$ on β. This suggests perhaps the $\sqrt{\beta}$ dependence is unavoidable, and it's curious if such a dependence can be tightened.*

[2]Our definition is slightly different from that of Suk and Kpotufe (2022); all mentioned results hold for either notions.

[3]From here on, we'll conflate the intervals $[a, b], [a, b)$ for $a, b \in \mathbb{N}$ with the natural numbers contained within.

5 Gap-Dependent Dynamic Regret Bounds

We next turn to the task of studying gap-dependent regret rates for non-stationary bandits. A first idea is to characterize the rate as that achieved by a *restarting oracle*, or an oracle procedure which restarts a stationary procedure at each changepoint. In other words, the gap-dependent dynamic regret rate is defined as the sum over stationary periods of the best possible stationary rates:

$$\sum_{\ell=1}^{L} \sum_{a:\delta_\ell(a)>0} \frac{\log(T)}{\delta_\ell(a)}, \tag{3}$$

where L is the number of stationary phases, the ℓ-th of which has gap profile $\{\delta_\ell(a)\}_{a\in[K]}$. Unfortunately, it's long been known in the switching bandit literature that, in the worst case, (3) cannot be attained simultaneously for different values of L (Garivier and Moulines, 2011; Lattimore and Szepesvári, 2020). The reason for such a hardness is intuitively because of the additional exploration required to detect unknown changes, which forces $\sqrt{T}$ regret.

Thus, a natural question remains: under what conditions can the rate (3) be achieved? Yet, before answering this, an even more basic question is glaringly unaddressed. Earlier (Section 1), we discussed alternative measures of non-stationarity (Besbes et al., 2019; Suk and Kpotufe, 2022; Abbasi-Yadkori et al., 2023), suggesting it is questionable whether (3) is even a sensible notion. For instance, (3) must scale with the number of stationary periods L which can be as large as T even while the total variation remains small (Besbes et al., 2019) or while there are no changes in best arm or significant shifts (Suk and Kpotufe, 2022; Abbasi-Yadkori et al., 2023). Thus, it remains to be seen if there is a better gap-dependent rate, which is invariant of irrelevant non-stationarity, which can also be achieved adaptively without knowledge of non-stationarity.

In our next contribution, we give answers to both these questions in terms of the significant shift oracle, introduced in Section 4. Recalling such an oracle roughly plays arms until they incur significant regret (1), and restarts at each significant shift, we'll see that a careful regret analysis of this oracle gives rise to a faster rate than (3) which is achievable adaptively in all environments without a significant shift.

Before getting into this, we summarize some of previous results in these directions.

5.1 Related Work on Gap-Dependent Regret

To start, we give an account of some works which aim to achieve the restarting oracle rate (3) under structured non-stationarity:

- Mukherjee and Maillard (2019) show a bound similar to (3) (albeit with a multiplicative factor which further depends on the difficulty of changes in gaps) under several assumptions on the changes: i.e., rewards of all arms change simultaneously, and changes are well-separated in time and large enough in magnitude so as to allow for fast-enough detection.
- Seznec et al. (2020) achieve the rate (3) in (restless) rotting bandits where rewards decrease in time.
- Besson et al. (2022) also study structured non-stationarity (that changes are delayed in time long enough for detection); they show a $\sqrt{LKT}$ regret bound on non-stationary instances where the minimum gap is $\Omega(1)$ and speculate, based on experimental findings, that their procedure could achieve faster logarithmic regret (as in (3)) on some problem instances.

To contrast, rather than directly making assumptions about the nature of changes, we show (Theorem 12) the restarting oracle rate (3) can be attained under any non-stationarity so long as a safe arm remains intact (a so-called *safe environment*; cf. Definition 10). In particular, changes of any kind (violating the structural assumptions listed above) are allowed in a safe environment. In general, however, we note our safe environment assumption is incomparable to the assumptions on changes made above.

On the other hand, we achieve rates *much faster than (3)* on safe environments, which is free of irrelevant non-stationarity (such as scaling with the raw number L of changes). The only other result, to our knowledge, which studies faster rates of this kind is Krishnamurthy and Gopalan (2021). For $K = 2$ armed bandits, they give an alternative gap-dependent rate in terms of a so-called *detectable gap profile* which quantifies what size aggregate gap is detectable over time (regardless of non-stationarity). However, while their proposed regret rate is logarithmic in the best case, it could scale like $\sqrt{T}$ even in safe environments. Furthermore, the only procedure given in Krishnamurthy and Gopalan (2021) achieving their gap-dependent rate requires some knowledge of non-stationarity.

5.2 Refined Regret Analysis of the Significant Shift Oracle

From the discussion of Appendix A of Suk and Kpotufe (2022), it is already evident that the significant shift oracle, which knows when arms incur significant regret, can attain safe regret of order $\sqrt{KT}$ on each significant phase. Here, we show a tighter gap-dependent regret rate which will arise from doing a more careful version of the analysis of Lemma 3 in Suk and Kpotufe (2022). To do so, we first set up some notation.

Notation 7. *Let $\mathcal{H}_t$ be be the σ-algebra generated by the random reward variables $\{Y_s(a)\}_{s \le t, a \in [K]}$ and exogenous time-varying randomness $\{\pi_s\}_{s \le t}$, as used by an algorithm π.*

We'll use $t_1, \ldots, t_K$ to denote ordered stopping times with respect to the filtration $\{\mathcal{H}_t\}_{t \in [T]}$ and use $\mathcal{S}_1, \cdots, \mathcal{S}_T$ to denote random subsets of $[K]$ which are adapted to this filtration.

Definition 8. *Let $t_0 \doteq 1$ and let $\mathcal{S}_1 \doteq [K]$. Then, we'll recursively define t_i and $\mathcal{S}_t$ for $t > t_{i-1}$ as follows: a stopping time $t_i > t_{i-1}$ is called an* **eviction time** *w.r.t. initial time t_{i-1} if $\mathcal{S}_{t_i - 1} = \mathcal{S}_{t_i - 2} = \cdots = \mathcal{S}_{t_{i-1}}$ and*

$$\forall a \in \mathcal{S}_{t_i - 1}, [s_1, s_2] \subseteq [1, t_i - 1] : \sum_{s=s_1}^{s_2} \frac{\delta_s(a)}{|\mathcal{S}_s|} \le C_2 \sqrt{\sum_{s=s_1}^{s_2} \frac{\log(T)}{|\mathcal{S}_s|}}. \tag{4}$$

We call $(t_1, t_2, \ldots, t_K)$ a **sequence of eviction times** *with associated* **safe armsets** *$\mathcal{S}_1 \supseteq \mathcal{S}_2 \supseteq \cdots \supseteq \mathcal{S}_T$.*

Remark 3. *Definition 8 can be seen as a generalization of Definition 4. The eviction time t_i serves as a more refined version of the first round when an arm becomes unsafe in the sense of (1) in Definition 4. The only major difference is that (4) more carefully involves the variance of estimating each arm's reward while uniformly exploring actions in the safe armsets $\mathcal{S}_t$ (as the significant shift oracle does). This modification is crucial for capturing the exact dependence on the number of arms when comparing to the restarting oracle rate (Theorem 9) and avoiding an extraneous $\log(K)$ factor in the analysis of Suk and Kpotufe (2022).*

Then, given Definition 8, we propose the following new gap-dependent rate

$$\mathscr{R}(\{t_i\}_{i \in [K]}, \{\mathcal{S}_t\}_{t \in [T]}, \pi) \doteq \sum_{i=1}^{K} \sum_{t=t_{i-1}}^{t_i - 1} \mathbb{E}_{a \sim \mathrm{Unif}\{\mathcal{S}_t\}}[\delta_t(a)], \tag{5}$$

Plainly speaking, (5) captures the regret of an elimination procedure which tracks the safe armsets $\mathcal{S}_t$ and uniformly explores $\mathcal{S}_t$ at ruond t. Note that (5) is a random quantity as the $t_i, \mathcal{S}_t$ may depend on the random rewards and exogenous randomness of some algorithm π. In fact, we'll next show that for any valid $t_i, \mathcal{S}_t$ satisfying Definition 8, the rate of (5) recovers (3) and is achievable adaptively in safe environments.

5.3 Properties of New Gap-Dependent Regret Rate

Proofs are deferred to Appendix C.

Theorem 9 (Recovering Restarting Oracle Rate)**.** *Let $\{t_i\}_{i \in [K], t \in [T]}$ be a sequence of eviction times and safe armsets per Definition 8. Then, for any environment with L stationary phases with the ℓ-th phase having* *gaps* $\delta_\ell(a)$ *for arm* a*, we have for any algorithm* $\{\pi_t\}_{t\in[T]}$*:*

$$\mathscr{R}(\{t_i\}_{i\in[K]}, \{\mathcal{S}_t\}_{t\in[T]}, \pi) \leq C_2^2 \sum_{\ell=1}^{L} \sum_{a:\delta_\ell(a)>0} \frac{\log(T)}{\delta_\ell(a)}.$$

We next show (5) recovers the usual $\sqrt{KT}$ regret bound when there's no significant shift.

Definition 10. *A bandit environment is called* **safe** *if, for any shrinking sequence of armsets* $\mathcal{G}_1 \supseteq \cdots \supseteq \mathcal{G}_T$*, there exists a* **safe arm** $a^\sharp$ *such that*

$$\forall [s_1, s_2] \subseteq [1, T] : \sum_{t=s_1}^{s_2} \frac{\delta_t(a^\sharp)}{|\mathcal{G}_t|} \leq C_3 \sqrt{\sum_{t=s_1}^{s_2} \frac{\log(T)}{|\mathcal{G}_t|}}. \tag{6}$$

In such an environment, any eviction times $\{t_i\}_{i=1}^K$ *and safe armsets* $\{\mathcal{S}_i\}_{t=1}^T$ *can be assumed WLOG to satisfy* $a^\sharp \in \mathcal{S}_T$ *and* $t_K = T + 1$*.*

Remark 4. *A safe environment is roughly one in which a significant shift is not permitted. This is because taking* $\mathcal{G}_t \equiv [K]$ *in (6) gives us* $\sum_{t=s_1}^{s_2} \delta_t(a^\sharp) < C_3\sqrt{(s_2 - s_1 + 1) \cdot K \log(T)}$ *which is a generalization of (1).*

Theorem 11 (Recovering $\sqrt{KT}$ Rate)**.** *We have, for any safe environment with eviction times and safe armsets* $\{t_i\}_{i\in[K],t\in[T]}$*, and algorithm* $\{\pi_t\}_{t\in[T]}$*:*

$$\mathscr{R}(\{t_i\}_{i\in[K]}, \{\mathcal{S}_t\}_{t\in[T]}, \pi) \leq C_2 \sqrt{KT \log(T)}.$$

5.4 Elimination Achieves Gap-Dependent Regret Rate

As already hinted up to this point, we posit that a randomized variant of elimination (similar to that of Suk and Kpotufe (2022)) in fact attains the rate of (5) in safe environments. We emphasize our contribution here is not algorithmic, but rather a tighter regret analysis of this algorithm to achieve the rate of (5).

Randomized Successive Elimination. We present a slightly different version of Algorithm 2 of Suk and Kpotufe (2022). First, we estimate the relative gap $\delta_t(a', a)$ via[4]:

$$\hat{\delta}_t(a', a) \doteq Y_t(a') \cdot \mathbf{1}\{\pi_t = a'\} - Y_t(a) \cdot \mathbf{1}\{\pi_t = a\}. \tag{7}$$

Then, at round t, we'll eliminate arms from an *active armset* $\mathcal{A}_t$ when an empirical analogue of (4) based on $\hat{\delta}_t(a', a)$ and $\mathcal{A}_t$ holds.

Algorithm 1: Randomized Successive Elimination

Initialize: $\mathcal{A}_t \leftarrow [K]$.
for $t = 1, 2, \ldots, T$ **do**
  Play a random arm $a \in \mathcal{A}_t$ selected with probability $1/|\mathcal{A}_t|$.
  Evict bad arms:
    $\mathcal{A}_t \leftarrow \mathcal{A}_t \backslash \left\{ a \in [K] : \exists \text{ round } t_0 \leq t \text{ s.t. } \max_{a' \in [K]} \sum_{s=t_0}^{t} \hat{\delta}_s(a', a) > C_5 \sqrt{\sum_{s=t_0}^{t} \frac{\log(T)}{|\mathcal{A}_s|}} \text{ holds} \right\}.$

Theorem 12. *Given any safe bandit environment over* T *rounds, letting* π *be Algorithm 1, we have w.p. at least* $1 - 1/T^2$*, for some algorithm-dependent eviction times and safe armsets* $\{t_i\}_{i\in[K]}, \{\mathcal{S}_t\}_{t\in[T]}$*:*

$$\sum_{t=1}^{T} \delta_t(\pi_t) \leq C_4 \cdot \left(\log(T) + \mathscr{R}(\{t_i\}_{i\in[K]}, \{\mathcal{S}_t\}_{t\in[T]}, \pi)\right).$$

[4] This estimator is slightly different than the importance-weighted one used in Suk and Kpotufe (2022); we require this estimate to mimic the form of (4).

Putting the previous results together, we conclude that elimination not only attains the gap-dependent regret of the restarting oracle $\sum_{\ell=1}^{L} \sum_{a:\delta_\ell(a)>0} \frac{\log(T)}{\delta_\ell(a)}$ but further attains a much faster rate $\mathscr{R}(\{t_i\}_{i\in[K]}, \{\mathcal{S}_t\}_{t\in[T]})$ which is free of irrelevant non-stationarity. In particular:

- There is no dependence in (5) on L, the number of changes in rewards or even the number S of best arm switches. We can in fact have $S, L = \Omega(T)$ while $\mathscr{R}(\{t_i\}_{i\in[K]}, \{\mathcal{S}_t\}_{t\in[T]}\}$ is small.
- As (5) only depends on the gaps, it is completely free of any changes in mean rewards which do not change the gaps.

On the other hand, as mentioned earlier, it's known in switching bandits that the restarting oracle's gap-dependent rate (3) cannot be achieved adaptively for unknown L. However, this does not contradict our findings because the constructed hard environment (e.g. Lattimore and Szepesvári, 2020, Theorem 31.2) is not safe (Definition 10). Thus, similar to Suk and Kpotufe (2022), we find that the notion of significant shift (which determines if an environment is safe or not) characterizes *difficult non-stationarity*, and so long as such a shift does not occur, we can attain the faster gap-dependent regret (5).

5.5 Lower Bound for Gap-Dependent Regret Rate

We next give a sense in which our new gap-dependent regret rate (5) is the best achievable rate. We do this by showing that the minimax regret rate over the class of all non-stationary environments with bounded $\mathscr{R}(\{t_i\}_{i\in[K]}, \{\mathcal{S}_t\}_{t\in[T]}, \pi) \leq R$ is $\Omega(R)$.

Theorem 13. *Let $\{t_i\}_{i=1}^K$ be an arbitrary set of rounds such that $t_{i+1} - t_i + 1 \geq K$ for all $i \in [K]$ with the convention that $t_0 \doteq 1$ and $t_{K-1} = t_K \doteq T+1$. Fix a positive real number R such that $R \leq \sum_{i=1}^{K-1} \sqrt{(t_i - t_{i-1}) \cdot (K+1-i)}$. Let $\mathscr{E}$ be the class of environments such that (a) $t_1, \ldots, t_K$ are valid deterministic eviction times with $C_2 = 1$ in (4) for some shrinking sequence of armsets $\mathcal{S}_1 \supseteq \cdots \supseteq \mathcal{S}_T$ and (b) such that:*

$$\sum_{i=1}^{K} \sum_{t=t_{i-1}}^{t_i - 1} \mathbb{E}_{a\sim \mathrm{Unif}\{\mathcal{S}_t\}}[\delta_t(a)] \leq R.$$

Then, for any algorithm π, we have:

$$\sup_{\mathcal{E}\in\mathscr{E}} R_{\mathcal{E}}(\pi, T) \geq \frac{1}{32 e^{25/12}} \cdot R.$$

Remark 5. *The reason Theorem 13 only considers the regime $R \leq \sum_{i=1}^{K-1} \sqrt{(t_{i+1} - t_i) \cdot (K+1-i)}$ is because (4) with $C_2 = 1$ implies for any valid eviction times $t_1, \ldots, t_K$ with $|\mathcal{S}_{t_i}| = K+1-i$:*

$$\sum_{t=t_{i-1}}^{t_i - 1} \sum_{a\in\mathcal{S}_{t_i}} \frac{\delta_t(a)}{K+1-i} \leq \sqrt{(t_i - t_{i-1}) \cdot (K+1-i)}.$$

Thus, environments with $\sum_{i=1}^{K-1} \sum_{t=t_{i-1}}^{t_i - 1} \mathbb{E}_{a\sim\mathrm{Unif}\{\mathcal{S}_t\}}[\delta_t(a)] > \sum_{i=1}^{K-1} \sqrt{(t_{i+1} - t_i) \cdot (K+1-i)}$ are not realizable in the class $\mathscr{E}$.

6 Achievability of Gap-Dependent Rate in terms of Smoothness

We've seen that the achievability of (5) hinges on whether an environment is safe (Definition 10), or roughly whether a significant shift occurs. In the smooth bandit model, a safe environment can be cleanly characterized via the value of the maximum Lipschitz constant or Hölder coefficient of the (normalized) gap function $f_a(x) \doteq \delta_{x\cdot T}(a)$ for arm a and all its derivatives:

$$\lambda_{\max} \doteq \max_{n=0,\ldots,\lfloor\beta\rfloor} \sup_{a\in[K]} \sup_{x\in[0,1]} |f_a^{(n)}(x)|.$$

In particular, an environment is safe if $\lambda_{\max} \le \sqrt{K/T}$, and so $\sqrt{K/T}$ is the critical value characterizing a phase transition in the achievable rates. The following result, whose proof mostly re-summarizes results shown in earlier sections, captures this phase transition.

Theorem 14 (Proof in Appendix C.5). *We have:*

(i) Any environment with $\lambda_{\max} \le \sqrt{K/T}$ *is safe.*

(ii) The minimax regret over the class of environments with $\lambda_{\max} > \sqrt{K/T}$ *is* $\Omega(\sqrt{KT})$.

References

Yasin Abbasi-Yadkori, András György, and Nevena Lazić. A new look at dynamic regret for non-stationary stochastic bandits. *Journal of Machine Learning Research*, 24(288):1–37, 2023. URL https://jmlr.org/papers/volume24/22-0387/22-0387.pdf.

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. URL https://arxiv.org/pdf/1902.00819.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002. URL http://rob.schapire.net/papers/AuerCeFrSc01.pdf.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best arm with an unknown number of distribution changes. *14th European Workshop on Reinforcement Learning (EWRL)*, 2018. URL https://ewrl.wordpress.com/wp-content/uploads/2018/09/ewrl_14_2018_paper_28.pdf.

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. URL https://proceedings.mlr.press/v99/auer19a/auer19a.pdf.

Peter L Bartlett. Learning with a slowly changing distribution. In *Proceedings of the fifth annual workshop on Computational Learning Theory (COLT)*, 1992. URL https://dl.acm.org/doi/pdf/10.1145/130385.130412.

Rakesh D Barve and Philip M Long. On the complexity of learning from drifting distributions. *Information and Computation*, 138(2):170–193, 1997. URL https://www.phillong.info/publications/distdrift.pdf.

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. URL https://pubsonline.informs.org/doi/epdf/10.1287/stsy.2019.0033.

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. URL https://jmlr.org/papers/volume23/20-1384/20-1384.pdf.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. *AISTATS*, 2011. URL https://proceedings.mlr.press/v15/beygelzimer11a/beygelzimer11a.pdf.

Moise Blanchard, Steve Hanneke, and Patrick Jaillet. Adversarial rewards in universal learning for contextual bandits. *arXiv preprint arXiv:2302.07186*, 2023. URL https://arxiv.org/pdf/2302.07186.

Sébastien Bubeck and Nicoló Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1), 2012. URL https://arxiv.org/pdf/1204.5721.pdf.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207 – 232, 2024. URL https://doi.org/10.1214/23-AOS2341.

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. URL https://proceedings.mlr.press/v89/cao19a/cao19a.pdf.

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. URL https://proceedings.mlr.press/v99/chen19b/chen19b.pdf.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page I–521–I–529. JMLR.org, 2014. URL https://proceedings.mlr.press/v32/combes14.pdf.

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. URL https://hal.science/hal-00281392/document.

Avishek Ghosh, Abishek Sankararaman, Kannan Ramchandran, Tara Javidi, and Arya Mazumdar. Competing bandits in non-stationary matching markets. *IEEE Transactions on Information Theory*, 2022. URL https://ieeexplore.ieee.org/document/10387415.

Melody Y Guan and Heinrich Jiang. Nonparametric stochastic contextual bandits. *AAAI*, 2018. URL https://cdn.aaai.org/ojs/11749/11749-13-15277-1-2-20201228.pdf.

Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. Smoothness-adaptive contextual bandits. *Operations Research*, 70(6):3198–3216, 2022. URL https://arxiv.org/pdf/1910.09714.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002. URL https://link.springer.com/book/10.1007/b97848.

Steve Hanneke and Liu Yang. Statistical learning under nonstationary mixing processes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1678–1686. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/hanneke19a/hanneke19a.pdf.

David P Helmbold and Philip M Long. Tracking drifting concepts using random examples. In *Proceedings of the fourth annual workshop on Computational Learning Theory (COLT)*, 1991. URL https://dl.acm.org/doi/10.5555/114836.114839.

David P Helmbold and Philip M Long. Tracking drifting concepts by minimizing disagreements. *Machine learning*, 14(1):27–45, 1994. URL https://link.springer.com/content/pdf/10.1007/BF00993161.pdf.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. *Conference on Learning Theory*, 2020. URL https://arxiv.org/pdf/1909.02553.

S. Jia, Qian Xie, Nathan Kallus, and P. Frazier. Smooth non-stationary bandits. In *International Conference on Machine Learning*, 2023. URL https://dl.acm.org/doi/10.5555/3618408.3619017.

Levente Kocsis and Csaba Szepesvári. Discounted ucb. *2nd PASCAL Challenges Workshop*, 2006. URL https://www.lri.fr/~sebag/Slides/Venice/Kocsis.pdf.

Junpei Komiyama, Edouard Fouché, and Junya Honda. Finite-time analysis of globally nonstationary multi-armed bandits. *arXiv preprint: arXiv:2107.11419*, 2021. URL https://arxiv.org/pdf/2107.11419.pdf.

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. URL http://proceedings.mlr.press/v99/krishnamurthy19a/krishnamurthy19a.pdf.

Ramakrishnan Krishnamurthy and Aditya Gopalan. On slowly-varying non-stationary bandits. *arXiv preprint: arXiv:2110.12916*, 2021. URL https://arxiv.org/pdf/2110.12916.pdf.

Tor Lattimore and Csaba Szepesvári. *Bandit Algoritms*. Cambridge University Press, 2020. URL https://tor-lattimore.com/downloads/book/book.pdf.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. *NIPS*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/97d98119037c5b8a9663cb21fb8ebf47-Paper.pdf.

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. URL https://dl.acm.org/doi/pdf/10.5555/3504035.3504482.

Philip M Long. The complexity of learning according to two models of a drifting environment. In *Conference on Computational Learning Theory (COLT)*, 1998. URL https://dl.acm.org/doi/pdf/10.1145/279943.279968.

Mark G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547 – 2554, 1997. URL https://doi.org/10.1214/aos/1030741084.

Tyler Lu, Dávid Pál, and Martin Pál. Showing relevant ads via context multi-armed bandits. In *Proceedings of AISTATS*, 2009. URL https://david.palenica.com/papers/clicks/lipschitz-clicks.pdf.

Anne Gael Manegueu, Alexandra Carpentier, and Yi Yu. Generalized non-stationary bandits. *arXiv preprint: arXiv:2102.00725*, 2021. URL https://arxiv.org/pdf/2102.00725.pdf.

Alessio Mazzetto and Eli Upfal. An adaptive algorithm for learning with unknown distribution drift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=exiXmAfuDK.

Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change detection. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 442–450, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL https://proceedings.mlr.press/v31/mellor13a.html.

Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, 2012. URL https://arxiv.org/pdf/1205.4343.

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. URL https://openreview.net/pdf?id=Ske_J_SrjE.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013. URL https://arxiv.org/pdf/1110.6084.

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. URL https://jmlr.org/papers/volume17/13-210/13-210.pdf.

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. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-10/issue-1/Randomized-allocation-with-arm-elimination-in-a-bandit-problem-with/10.1214/15-EJS1104.full.

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. URL https://proceedings.mlr.press/v83/reeve18a/reeve18a.pdf.

Phillipe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *COLT*, 2010. URL https://arxiv.org/pdf/1003.1630.

Jyotirmoy Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, pages 1978–2002, 1991. URL https://projecteuclid.org/journals/annals-of-statistics/volume-19/issue-4/One-Armed-Bandit-Problems-with-Covariates/10.1214/aos/1176348382.full.

Igal Sason and Sergio Verdú. f -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. URL https://arxiv.org/pdf/1508.00335.

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2564–2572, 2019. URL https://proceedings.mlr.press/v89/seznec19a/seznec19a.pdf.

Julien Seznec, Pierre Menard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020. URL https://proceedings.mlr.press/v108/seznec20a/seznec20a.pdf.

Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014. URL https://arxiv.org/pdf/0907.3986.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. ISSN 1935-8237. doi: 10.1561/2200000068. URL https://arxiv.org/pdf/1904.07272.pdf.

Joe Suk and Samory Kpotufe. Self-tuning bandits over unknown covariate-shifts. *International Conference on Algorithmic Learning Theory (ALT)*, 2021. URL https://proceedings.mlr.press/v132/suk21a/suk21a.pdf.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. *Conference on Learning Theory (COLT)*, 2022. URL https://proceedings.mlr.press/v178/suk22a/suk22a.pdf.

Joe Suk and Samory Kpotufe. Tracking most significant shifts in nonparametric contextual bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/pdf/2307.05341.pdf.

Francesco Trovò, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *J. Artif. Intell. Res.*, 68:311–364, 2020. URL https://www.jair.org/index.php/jair/article/view/11407.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009. ISBN 978-0-387-79051-0. URL https://doi.org/10.1007/b13794.

Lai Wei and Vaihbav Srivatsva. On abruptly-changing and slowly-varying multiarmed bandit problems. *Annual American Control Conference (ACC)*, 2018. URL https://ieeexplore.ieee.org/document/8431265.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979. URL https://dept.stat.lsa.umich.edu/~michaelw/PPRS/1979jasa.pdf.

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. URL https://projecteuclid.org/journals/annals-of-statistics/volume-30/issue-1/Randomized-Allocation-with-nonparametric-estimation-for-amulti-armed-bandit-problem/10.1214/aos/1015362186.full.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. *Proceedings of the 26th Annual international Conference on Machine Learning*, pages 1177–1184, 2009. URL http://www.machinelearning.org/archive/icml2009/papers/367.pdf.

A Proof of Dynamic Regret Lower Bound (Theorem 3)

Overview of Argument. The construction will rely on bump reward functions which also appear in the classical minimax lower bounds for integrated risk of nonparametric regression of (β, L)-Hölder functions (e.g., Section 2.5 of Tsybakov, 2009). This will be combined with a Le Cam's method style of argument for establishing a regret lower bound for stationary bandits (Lattimore and Szepesvári, 2020, Theorem 15.2), which we'll modify to work for segments of mild non-stationarity.

Preliminaries. First, getting some trivial cases out of the way, let's assume that $K \leq T/4$ or else we can just trivially show a lower bound of order K using a stationary construction (which is always (β, λ)-Hölder for any $\beta, \lambda > 0$).

Let's also assume that

$$\sqrt{KT} \leq T^{\frac{\beta+1}{2\beta+1}} \cdot \lambda^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}} \iff \sqrt{K/T} \leq \lambda,$$

or else again we can appeal to the well-known $\sqrt{KT}$ stationary lower bound.

Now, let $\tilde{\lambda} \doteq (2^{-(2\beta+1)} \cdot (T/K)^{\beta}) \wedge \lambda$. Let $M \doteq \left\lceil T^{\frac{1}{2\beta+1}} \cdot K^{-\frac{1}{2\beta+1}} \cdot \tilde{\lambda}^{\frac{2}{2\beta+1}} \right\rceil$. Now, since $\tilde{\lambda} \leq (T/K)^{\beta} \cdot 2^{-(2\beta+1)}$, we have that $M \leq \lceil T/4 \rceil$. Next, we argue that WLOG M divides T. If this is not the case, then we can replace the horizon T with $T_0 \doteq M \cdot \lfloor T/M \rfloor \leq T$, which is a multiple of M, and show the lower bound for T_0 which suffices for the end result since $T_0 \geq T - M \geq T/2$.

At a high level, we'll construct M instances of a randomly selected (β, λ)-Hölder environment of length T/M. We'll then argue that the concatenation of any such realized M environments is itself a (β, λ)-Hölder environment over T rounds.

Intuitively, over each period of length T/M, the constructed sub-environment will be nearly stationary and ensure a regret lower bound of order $\sqrt{K \cdot (T/M)}$. Then, summing over the M sub-environments and taking a random prior over choice of instances, we get a dynamic regret lower bound of order

$$\sqrt{TKM} \geq T^{\frac{\beta+1}{2\beta+1}} \cdot \tilde{\lambda}^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}}.$$

If $\tilde{\lambda} = \lambda$, we are done. If $\tilde{\lambda} < \lambda$, then $T^{\frac{\beta+1}{2\beta+1}} \lambda^{\frac{1}{2\beta+1}} K^{\frac{\beta}{2\beta+1}} \geq c_0 T$ so that it suffices to show a linear regret lower bound. Since $\tilde{\lambda} < \lambda \implies \tilde{\lambda} \propto (T/K)^{\beta}$, plugging this into the above RHS indeed gives us said linear regret lower bound.

We proceed by first defining the sub-environment over T/M rounds.

Defining Bump Function Rewards. First, define the function $\varphi : [0,1] \to \mathbb{R}_{\geq 0}$ as:

$$\varphi(x) \doteq \frac{\tilde{\lambda} \cdot h^{\beta}}{2} \cdot \Phi\left(\frac{x - h/2}{h}\right),$$

where $h \doteq 1/M$ is a bandwidth and Φ is the C^{∞} bump function

$$\Phi(u) \doteq \exp\left(-\frac{1}{1-u^2}\right) \cdot \mathbf{1}\{\|u\| \leq 1\}.$$

Now, consider an *assignment* of M best arms $\mathbf{a} \doteq (a_1, \ldots, a_M)$. For an arm $a \in [K]$, we define the function $\varphi_{a,\mathbf{a},i}(x)$ as

$$\varphi_{a,\mathbf{a},i}(x) \doteq \begin{cases} \varphi(x) & a = a_i \\ -\varphi(x) & a \neq a_i \end{cases}.$$

Then, for assignment $\mathbf{a}$, the reward function of arm a will be defined as:

$$\mu_{t,\mathbf{a}}(a) \doteq \frac{1}{2} + \sum_{i=1}^{M} \varphi_{a,\mathbf{a},i}(t/T).$$

The above are valid bounded reward functions in $[0, 1]$ since by the definition of $\tilde{\lambda}$:

$$\frac{\tilde{\lambda} \cdot h^\beta}{2} \leq \frac{1}{2}.$$

Next, we claim that for any arm assignment $\mathbf{a}$, the induced bandit environment is (β, λ)-Hölder (Definition 2).

First, since φ is $(\beta, \lambda/2)$-Hölder (assertion (a) of Tsybakov, 2009, Section 2.5), the reward function $t \mapsto \mu_{t,\mathbf{a}}(a)$ for each arm a is $(\beta, \lambda/2)$-Hölder being a sum of $(\beta, \lambda/2)$-Hölder functions with disjoint supports. Then, the gap functions $t \mapsto \max_{a'} \mu_{t,\mathbf{a}}(a') - \mu_{t,\mathbf{a}}(a)$ are (β, λ)-Hölder as a difference of two $(\beta, \lambda/2)$-Hölder functions.

In fact, we note the induced environments would also be (β, λ)-Hölder if we had defined $\varphi_{a,\mathbf{a},i}$ as:

$$\tilde{\varphi}_{a,\mathbf{a},i}(x) \doteq \begin{cases} 0 & a = a_i \\ -\varphi(x) & a \neq a_i \end{cases}.$$

In what follows, we'll make use of environments which use both formulas $\varphi_{a,\mathbf{a},i}(x)$ and $\tilde{\varphi}_{a,\mathbf{a},i}(x)$.

Lower Bound for Sub-Environment. Letting $n \doteq T/M$, we will show a regret lower bound over a sub-environment of n rounds where there's a fixed optimal arm. In particular, we claim that, over a sub-environment of n rounds, the gap of any suboptimal arm will be $\Omega\left(\tilde{\lambda} \cdot (n/T)^\beta\right)$ over a subdomain of length $\Omega(n)$. This will be enough to sum up regret lower bounds over M different sub-environments.

Going into more detail, for $x \in [h/8, 7h/8]$, observe that

$$\varphi(x) \geq \frac{\tilde{\lambda} \cdot h^\beta}{2} \cdot \exp\left(-\frac{1}{1 - \left(\frac{3}{8}\right)^2}\right) \geq \frac{\tilde{\lambda} \cdot h^\beta}{10}. \tag{8}$$

Now, consider a sub-environment $\mathcal{E}_1$ over n rounds (which we'll for ease momentarily parametrize via $[n]$) on which arm 1 is optimal with:

$$\forall t \in [n] : \mu_t(a) \doteq \frac{1}{2} + \begin{cases} 0 & a = 1 \\ -\varphi(t/T) & a \neq 1 \end{cases}.$$

For any algorithm π, there must exist an arm $a \neq 1$ for which the arm-pull count $N_n(a) \doteq \sum_{t=1}^n \mathbf{1}\{\pi_t = a\}$ satisfies $\mathbb{E}_{\mathcal{E}_1}[N_n(a)] \leq \frac{n}{K-1}$ since $\sum_{a=2}^K \mathbb{E}_{\mathcal{E}_1}[N_n(a)] = n$. Now, consider the environment $\mathcal{E}_a$ on which arm a is instead optimal with reward function:

$$\forall t \in [n] : \mu_t(a) \doteq \frac{1}{2} + \varphi(t/T).$$

The reward functions of all arms other than a in $\mathcal{E}_a$ are defined identically to that of $\mathcal{E}_1$.

Then, if $N_n(1) \leq n/2$ in environment $\mathcal{E}_1$, then by pigeonhole at least $n/4$ rounds of the rounds in $[n/8, 7n/8]$ must consist of suboptimal arm pulls paying a per-round regret of at least $\Delta \doteq \frac{\tilde{\lambda} \cdot h^\beta}{10}$ by (8). Similarly, under environment $\mathcal{E}_a$, if arm 1 is pulled more than $n/2$ times, then at least $n/4$ of the rounds in $[n/8, 7n/8]$ must consist of pulls of arm 1 which forces a regret of at least Δ. Thus, we lower bound the total regret over n rounds in $\mathcal{E}_1$ and $\mathcal{E}_a$ by:

$$R_{\mathcal{E}_1}(\pi, n) \geq \frac{n\Delta}{4} \cdot \mathbb{P}_{\mathcal{E}_1}(N_n(1) \leq n/2)$$
$$R_{\mathcal{E}_a}(\pi, n) \geq \frac{n\Delta}{4} \cdot \mathbb{P}_{\mathcal{E}_a}(N_n(1) > n/2).$$

Then, combining the above two displays with the Bretagnolle-Huber inequality (Lemma 20), we have:

$$R_{\mathcal{E}_1}(\pi, n) + R_{\mathcal{E}_a}(\pi, n) \geq \frac{n\Delta}{4}(\mathbb{P}_{\mathcal{E}_1}(N_n(1) \leq n/2) + \mathbb{P}_{\mathcal{E}_a}(N_n(1) > n/2)) \geq \frac{n\Delta}{8} \exp\left(-\mathrm{KL}(\mathcal{P}_1, \mathcal{P}_a)\right), \tag{9}$$

where $\mathcal{P}_1, \mathcal{P}_a$ are the respective induced distributions on the history of observations and decisions over the n rounds. We next decompose this KL divergence using chain rule:

$$\mathrm{KL}(\mathcal{P}_1, \mathcal{P}_a) = \sum_{t=1}^{n} \mathbb{E}_{\mathcal{P}_1}[\mathbf{1}\{\pi_t = a\}] \cdot \mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2} - \varphi\left(\frac{t}{T}\right)\right), \mathrm{Ber}\left(\frac{1}{2} + \varphi\left(\frac{t}{T}\right)\right)\right).$$

Next, we bound the KL between Bernoulli's. Let $\Delta_t \doteq \varphi(t/T)$. Then:

$$\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2} - \varphi\left(\frac{t}{T}\right)\right), \mathrm{Ber}\left(\frac{1}{2} + \varphi\left(\frac{t}{T}\right)\right)\right) = \left(\frac{1}{2} + \Delta_t\right) \log\left(\frac{1/2 + \Delta_t}{1/2 - \Delta_t}\right) + \left(\frac{1}{2} - \Delta_t\right) \log\left(\frac{1/2 - \Delta_t}{1/2 + \Delta_t}\right).$$

Elementary calculations show the above RHS expression is at most $10\Delta_t^2$ for $\Delta_t \leq 1/4$, which holds for all $t \in [n]$ since by the definition of $\tilde{\lambda}$ and since $\Phi(x) \leq 1$ for all x:

$$\forall t \in [n] : \varphi(t/T) \leq \frac{\tilde{\lambda} \cdot h^\beta}{2} \leq \frac{1}{4}.$$

Recalling that $\Delta \doteq \frac{\tilde{\lambda} \cdot h^\beta}{10}$, the above can also be re-phrased as $\Delta_t \leq 5\Delta$ for $t \in [n]$. Thus, we obtain a KL bound of:

$$\mathrm{KL}(\mathcal{P}_1, \mathcal{P}_a) \leq \mathbb{E}_{\mathcal{E}_1}[N_n(a)] \cdot 250\Delta^2 \leq \frac{n}{K-1} \cdot 250 \cdot \Delta^2 \leq \frac{K}{K-1} \cdot 2.5,$$

where the last inequality follows from $\Delta \leq \frac{1}{10}\sqrt{\frac{K}{n}}$ (which holds from the definition of M). Noting that

$$\exp\left(-\frac{K}{K-1} \cdot 2.5\right) \geq \frac{1}{100}$$

for all $K \geq 2$, the sub-environment lower bound of order $\Omega(n \cdot \tilde{\lambda} \cdot (n/T)^\beta)$ is concluded by combining the above display with (9).

Concatenating Different Sub-Environments. We first claim that the pair of environments $\mathcal{E}_1$ and $\mathcal{E}_a$ can be analogously constructed for every length n interval of rounds $\{i \cdot M + 1, \ldots, (i+1) \cdot M\}$ for $i \in [T/M - 1]$. First, note that the expected dynamic regret can be written as:

$$R(\pi, T) = \mathbb{E}\left[\sum_{t=1}^{T} \delta_t(\pi_t)\right] = \sum_{i=0}^{T/M-1} \mathbb{E}\left[\mathbb{E}\left[\sum_{t=i \cdot M+1}^{(i+1) \cdot M} \delta_t(\pi_t) \mid \mathcal{H}_{i \cdot M}\right]\right],$$

where $\mathcal{H}_t$ is the filtration of history of observations and decisions up to round t. Now, for $i \in [T/M - 1]$, there must exist an arm $a \neq 1$ whose conditional arm-pull count $\mathbb{E}\left[\sum_{t=i \cdot M+1}^{(i+1) \cdot M} \mathbf{1}\{\pi_t = a\} \mid \mathcal{H}_{i \cdot M}\right] \leq \frac{n}{K-1}$. Then, conditional on $\mathcal{H}_{i \cdot M}$, we can design environments $\mathcal{E}_1$ and $\mathcal{E}_a$ as before and lower bound the conditional regret by $\Omega(n \cdot \tilde{\lambda} \cdot (n/T)^\beta)$.

Putting everything together, we have there exists an assignment $\mathbf{a} = (a_1, \ldots, a_M)$ of arms and a concatenated bandit environment consisting of bump function rewards as defined earlier for which the total regret is lower bounded by:

$$M \cdot n \cdot \tilde{\lambda} \cdot \left(\frac{n}{T}\right)^\beta \geq c_1 \min\{T^{\frac{1+\beta}{2\beta+1}} \cdot \tilde{\lambda}^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}}, T \cdot \tilde{\lambda}\} \propto T^{\frac{1+\beta}{2\beta+1}} \cdot \tilde{\lambda}^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}},$$

where the last equality holds from the definition of $\tilde{\lambda}$ and the earlier assumptions of $K \leq T/4$ and $\sqrt{K/T} \leq \lambda$ (which were made to rule out trivial cases).

B Proof of Dynamic Regret Upper Bound (Theorem 5)

Overview of Argument. The main idea is that on each significant phase $[\tau_i, \tau_{i+1},)$, any arm a which is at one point optimal, i.e. $\delta_t(a) = 0$ for some $t \in [\tau_i, \tau_{i+1})$, must have a gap function $t \mapsto \delta_t(a)$ with large variation within the phase since the gap must also at some point be large because of our notion of significant regret (see Fact 16). More generally, if the gap function is $\lfloor \beta \rfloor$ times differentiable, then we can find an order $\lfloor \beta \rfloor$ critical point of the gap function across $\lfloor \beta \rfloor$ different phases using Rolle's Theorem. Using the definition of Hölder function (Definition 1), we can then bound the derivatives of the gap functions using this critical point. Such bounds can in turn be plugged into an order-$\lfloor \beta \rfloor$ Taylor approximation of the gap function. Ultimately, these calculations allow us to relate the phase length $\tau_{i+1} - \tau_i$ to the smoothness parameters β, λ. The key claim is that each phase must be roughly at least length $T^{\frac{2\beta}{2\beta+1}} \lambda^{-\frac{2}{2\beta+1}} K^{\frac{1}{2\beta+1}}$ which gives the desired regret bound.

The actual proof will require a bit more care as the optimal arm can change every round within a phase and some phases may be too short to have sufficient variation.

Getting into the proof, we first establish two key facts about significant phases which will be crucial later on.

Fact 15. *Each significant phase $[\tau_i, \tau_{i+1})$ with $\tau_{i+1} \neq T + 1$ is length at least K.*

Proof. This is true by our notion of significant regret (1) since we must have for some arm a and interval $[s_1, s_2] \subseteq [\tau_i, \tau_{i+1})$:

$$s_2 - s_1 + 1 \geq \sum_{t=s_1}^{s_2} \delta_t(a) \geq \sqrt{K \cdot (s_2 - s_1 + 1)} \implies s_2 - s_1 + 1 \geq K.$$

Thus, $\tau_{i+1} - \tau_i \geq s_2 - s_1 + 1 \geq K$. ■

Fact 16. *For each significant phase $[\tau_i, \tau_{i+1})$ with $\tau_{i+1} \neq T + 1$, we must have for each arm $a \in [K]$, there exists a round $t \in [\tau_i, \tau_{i+1}]$ such that:*

$$\delta_t(a) \geq \sqrt{\frac{K}{\tau_{i+1} - \tau_i + 1}}.$$

Proof. By Definition 4, we have for each arm $a \in [K]$, there exists $[s_1, s_2] \subseteq [\tau_i, \tau_{i+1}]$ such that:

$$\sum_{t=s_1}^{s_2} \delta_t(a) \geq \sqrt{K \cdot (s_2 - s_1 + 1)} \geq \sum_{t=s_1}^{s_2} \sqrt{\frac{K}{\tau_{i+1} - \tau_i + 1}}.$$

The conclusion follows. ■

Now, if $T \leq K$, then the desired regret rate is vacuous so we are done. Suppose $T > K$.

Let $m \doteq \lfloor \beta \rfloor$. We next decompose the dynamic regret bound in terms of significant phases according to the length of the significant phase:

$$\sum_{i=0}^{\tilde{L}} \sqrt{K \cdot (\tau_{i+1} - \tau_i)} = \sqrt{KT} + \sum_{i:\tau_{i+1}-\tau_i > (m+1)\cdot K} \sqrt{K \cdot (\tau_{i+1} - \tau_i)} + \sum_{i:\tau_{i+1}-\tau_i \leq (m+1)\cdot K} \sqrt{K \cdot (\tau_{i+1} - \tau_i)},$$

Our goal will be to show each of the two sums on the RHS above are of order

$$\sqrt{m+1}\left(\sqrt{KT} + T^{\frac{\beta+1}{2\beta+1}} \cdot \lambda^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}}\right).$$

Note that, going forward in the proof, we'll constrain our attention to significant phases $[\tau_i, \tau_{i+1})$ such that $\tau_{i+1} \neq T + 1$ as the $\sqrt{KT}$ term on the above RHS accounts for this regret contributed by this final phase.

Bounding Regret over Long Phases. We first bound the regret over long significant phases $[\tau_i, \tau_{i+1})$ such that $\tau_{i+1} - \tau_i > (m+1) \cdot K$. By the pigeonhole principle, there must exist an arm $a \in [K]$ which is optimal (i.e., $a \in \operatorname{argmax}_{a \in [K]} \mu_t(a)$) for at least $(m+1)$ different rounds in $[\tau_i, \tau_{i+1})$. Fix such an arm a. By Definition 1, there exists a (β, λ)-Hölder interpolating function $F : [0,1] \to [0,1]$ such that $F(t/T) = \delta_t(a)$. By the definition of a, there exists at least $(m+1)$ different rounds $t \in [\tau_i, \tau_{i+1})$ such that $F(t/T) = 0$. By Rolle's Theorem, this means there exists a point $x_0 \in [\tau_i/T, \tau_{i+1}/T)$ such that $F^{(m)}(x_0) = 0$.

Then, by the definition of a (β, λ)-Hölder function (Definition 1), note that

$$\forall x \in \left[\frac{\tau_i}{T}, \frac{\tau_{i+1}}{T}\right] : |F^{(m)}(x)| = |F^{(m)}(x) - F^{(m)}(x_0)| \leq \lambda \cdot \left(\frac{\tau_{i+1} - \tau_i}{T}\right)^{\beta - m}, \tag{10}$$

First, suppose $m = 0$. Then, we have by Fact 16 that $F(x) = \delta_{x \cdot T}(a) \geq \sqrt{\frac{K}{\tau_{i+1} - \tau_i + 1}}$ for some $x \in [\tau_i/T, \tau_{i+1}/T]$. Combining this with our above display, there exists $x \in [\tau_i/T, \tau_{i+1}/T]$ such that for $m = 0$:

$$\sqrt{\frac{K}{\tau_{i+1} - \tau_i + 1}} \leq F(x) = |F^{(m)}(x)| \leq \lambda \cdot \left(\frac{\tau_{i+1} - \tau_i + 1}{T}\right)^{\beta - m}.$$

We will next argue, using a Taylor approximation, that the above inequalities also essentially hold for $m \geq 1$.

If $m \geq 1$, then by the Mean Value Theorem and (10):

$$\begin{aligned}\forall x \in \left[\frac{\tau_i}{T}, \frac{\tau_{i+1}}{T}\right] : |F^{(m-1)}(x)| &\leq \sup_{x'} |F^{(m)}(x')| \sup_{y,z \in \left[\frac{\tau_i}{T}, \frac{\tau_{i+1}}{T}\right]} |y - z| \\ &\leq \lambda \cdot \left(\frac{\tau_{i+1} - \tau_i}{T}\right)^{\beta - m} \left(\frac{\tau_{i+1} - \tau_i}{T}\right).\end{aligned}$$

Then, by induction and repeatedly applying the Mean Value Theorem, we have:

$$\forall k \in \{0, \ldots, m-1\} : |F^{(k)}(x)| \leq \lambda \cdot \left(\frac{\tau_{i+1} - \tau_i}{T}\right)^{\beta - k}. \tag{11}$$

Then, taking an order-$(m-1)$ Taylor expansion with Lagrange remainder of F about a root $x_1 \in [\tau_i/T, \tau_{i+1}/T)$ (which we already argued exists since a is optimal at some round in $[\tau_i, \tau_{i+1})$), we have there exists $\xi \in [\tau_i/T, \tau_{i+1}/T)$ such that:

$$\begin{aligned}\forall x \in \left[\frac{\tau_i}{T}, \frac{\tau_{i+1}}{T}\right) : |F(x)| &= |F(x) - F(x_1)| && (F(x_1) = 0) \\ &\leq \sum_{k=1}^{m-1} \frac{|F^{(k)}(x_1)|}{k!} \cdot |x - x_1|^k + \frac{|F^{(m)}(\xi)|}{m!} \cdot |x - x_1|^m && \text{(Taylor's Theorem)} \\ &\leq (e - 1) \cdot \lambda \cdot \left(\frac{\tau_{i+1} - \tau_i}{T}\right)^{\beta} && \text{(from (11))}.\end{aligned}$$

Now, as before, there must exist an $x \in [\tau_i/T, \tau_{i+1}/T)$ such that (using the above display):

$$\sqrt{\frac{K}{\tau_{i+1} - \tau_i + 1}} \leq F(x) \leq (e - 1) \cdot \lambda \cdot \left(\frac{\tau_{i+1} - \tau_i + 1}{T}\right)^{\beta}.$$

Rearranging, the above implies

$$\tau_{i+1} - \tau_i + 1 \geq (e - 1)^{-\frac{2}{2\beta+1}} \cdot T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}.$$

Now, letting M be the total number of long significant phases, we must have

$$2T \geq T + \tilde{L} \geq \sum_{i < \tilde{L} : \tau_{i+1} - \tau_i > (m+1) \cdot K} \tau_{i+1} - \tau_i + 1 \geq M \cdot (e - 1)^{-\frac{2}{2\beta+1}} \cdot T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}.$$

Thus,

$$M \leq (e-1)^{\frac{2}{2\beta+1}} \frac{2T}{T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}}.$$

Then, we have by Jensen's inequality:

$$\begin{aligned}\sum_{i<\tilde{L}:\tau_{i+1}-\tau_i>(m+1)\cdot K} \sqrt{K \cdot (\tau_{i+1}-\tau_i)} &\leq \sqrt{K \cdot T \cdot M} \\ &\leq (e-1)^{\frac{1}{2\beta+1}} \cdot \sqrt{K \cdot T \cdot \left(\frac{2T}{T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}}\right)} \\ &= \sqrt{2}(e-1)^{\frac{1}{2\beta+1}} \cdot T^{\frac{\beta+1}{2\beta+1}} \cdot \lambda^{\frac{1}{2\beta+1}} \cdot K^{\frac{\beta}{2\beta+1}}.\end{aligned}$$

Bounding Regret over Short Phases. We next analyze the short significant phases $[\tau_i, \tau_{i+1})$ where $\tau_{i+1} - \tau_i \leq (m+1) \cdot K$. The difficulty here is that we cannot directly apply the same argument as we did for long phases since there may not exist $m+1$ different rounds where an arm is optimal within the phase. To get around this, we'll concatenate different short phases together and construct *pseudo phases* where we can apply the argument as we made for long phases.

Definition 17. *Let n_0 be the smallest significant shift τ_i belonging to a short significant phase $[\tau_i, \tau_{i+1})$. Then, recursively define n_{j+1} to be the smallest significant shift $\tau_{i+1} > n_j$ corresponding to a short significant phase $[\tau_i, \tau_{i+1})$ such that*

$$[n_j, \tau_{i+1}) \subseteq \bigcup_{i<\tilde{L}:\tau_{i+1}-\tau_i \leq (m+1)\cdot K} [\tau_i, \tau_{i+1}),$$

and such that $\tau_{i+1} - n_j \geq (m+1) \cdot K$. If no such significant shift τ_{i+1} exists, let n_{j+1} be the largest significant shift τ_{i+1} such that

$$[n_j, \tau_{i+1}) \subseteq \bigcup_{i<\tilde{L}:\tau_{i+1}-\tau_i \leq (m+1)\cdot K} [\tau_i, \tau_{i+1}).$$

We call $[n_j, n_{j+1})$ a **pseudo phase**. *The sequence $n_0, n_1, \ldots$ induces a partition of short phases:*

$$\bigsqcup_{i<\tilde{L}:\tau_{i+1}-\tau_i \leq (m+1)\cdot K} [\tau_i, \tau_{i+1}) = \bigsqcup_j [n_j, n_{j+1}).$$

Call a pseudo phase $[n_j, n_{j+1})$ **filled** *if $n_{j+1} - n_j \geq (m+1) \cdot K$, and* **unfilled** *otherwise.*

Intuitively, a filled pseudo phase is sufficiently long and will be of similar length to a long phase.

We'll now further decompose the dynamic regret over short phases using Jensen's inequality and the fact that each pseudo phase $[n_j, n_{j+1})$ can contain at most $(m+1)$ short phases $[\tau_i, \tau_{i+1})$ since all phases are length at least K (Fact 15):

$$\sum_{i<\tilde{L}:\tau_{i+1}-\tau_i \leq (m+1)\cdot K} \sqrt{K \cdot (\tau_{i+1}-\tau_i)} \leq \sum_j \sqrt{K \cdot (n_{j+1}-n_j) \cdot (m+1)}.$$

Next, we further decompose the above RHS sum over pseudo phases into sums over filled and unfilled pseudo phases. Thus, using Jensen again, it suffices to bound

$$\sqrt{(m+1) \cdot K \cdot T \cdot J_1} + \sqrt{(m+1) \cdot K \cdot T \cdot J_2}, \tag{12}$$

where J_1 and J_2 are respectively the number of filled and unfilled pseudo phases. We'll first bound J_1.

Bounding the Number of Filled Pseudo Phases. This will proceed similarly to the argument for long phases. Fix a filled pseudo phase $[n_j, n_{j+1})$ which has $n_{j+1} - n_j \geq (m+1) \cdot K$. Then, by the pigeonhole principle, there must exist an arm $a \in [K]$ which is optimal in at least $(m+1)$ different rounds in $[n_j, n_{j+1})$. Then, using the same arguments as before except replacing the long phase $[\tau_i, \tau_{i+1})$ with the pseudo phase $[n_j, n_{j+1})$, we conclude that there exists $x \in [n_j/T, n_{j+1}/T)$ for which:

$$\sqrt{\frac{K}{n_{j+1} - n_j + 1}} \leq \delta_{x \cdot T}(a) \leq (e-1) \cdot \lambda \cdot \left(\frac{n_{j+1} - n_j + 1}{T} \right)^{\beta}.$$

Rearranging, we get

$$n_{j+1} - n_j + 1 \geq (e-1)^{-\frac{2}{2\beta+1}} \cdot T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}.$$

Then, via similar arguments to before, the number of filled pseudo phases is at most

$$J_1 \leq (e-1)^{\frac{2}{2\beta+1}} \frac{2T}{T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}}.$$

Plugging this into (12) gives the desired regret bound for $\sqrt{KTJ_1}$.

Bounding the Number of Unfilled Pseudo Phases. Since unfilled pseudo phases are not of sufficient length $n_{j+1} - n_j < (m+1) \cdot K$, further care is required to make use of Rolle's Theorem. The key workaround is that each unfilled pseudo phase can be extended into an interval of length at least $(m+1) \cdot K$ without overcounting rounds, essentially because the unfilled pseudo phases are well-separated in time by Definition 17.

First, handling an edge case, suppose there are no long phases $[\tau_i, \tau_{i+1})$ with $\tau_{i+1} - \tau_i \geq (m+1) \cdot K$ and no filled pseudo phases $[n_j, n_{j+1})$ with $n_{j+1} - n_j \geq (m+1) \cdot K$. By Definition 17, this means there is just one pseudo phase $[n_j, n_{j+1})$ which subsumes all the significant phases $[\tau_i, \tau_{i+1})$ with $\tau_{i+1} \neq T+1$. Thus, in this case, $J_2 = 1$ and we are done.

Now, suppose there are at least two unfilled pseudo phases. Then, by Definition 17, two consecutive unfilled pseudo phases must be separated by at least one long phase $[\tau_i, \tau_{i+1})$. Let $I_1, I_2, \ldots$ be the unfilled pseudo phases ordered by start times. Then, each $I_j = [n_{j'}, n_{j'+1})$ has a posterior long phase $[\tau_i, \tau_{i+1})$ such that $\tau_i = n_{j'+1}$ and $\tau_{i+1} - n_{j'} \geq \tau_{i+1} - \tau_i \geq (m+1) \cdot K$.

Then, applying the same chain of reasoning as before to the interval $[n_{j'}, \tau_{i+1})$, we conclude there exists an arm $a \in [K]$ and $x \in [n_{j'}/T, \tau_{i+1}/T)$ for which:

$$\sqrt{\frac{K}{\tau_{i+1} - n_{j'} + 1}} \leq \delta_{x \cdot T}(a) \leq (e-1) \cdot \lambda \cdot \left(\frac{\tau_{i+1} - n_{j'} + 1}{T} \right)^{\beta} \implies$$
$$\tau_{i+1} - n_{j'} + 1 \geq (e-1)^{-\frac{2}{2\beta+1}} \cdot T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}.$$

Now, for each unfilled pseudo phase $I_j = [n_{j'}, n_{j'+1},)$, let $\overline{I}_j$ be the extension to the posterior long phase $[n_{j'}, \tau_{i+1})$ per our previous discussion. Then, since the $\overline{I}_j$ are mutually disjoint, we have the number of unfilled pseudo phases J_2 is at most one greater than the number of extended intervals $\overline{I}_j$. Thus, via similar arguments to before:

$$J_2 \leq 1 + (e-1)^{\frac{2}{2\beta+1}} \frac{2T}{T^{\frac{2\beta}{2\beta+1}} \cdot \lambda^{-\frac{2}{2\beta+1}} \cdot K^{\frac{1}{2\beta+1}}}.$$

As before, plugging this into (12) give the desired regret bound for $\sqrt{KTJ_2}$.

C Proofs of Results about Gap-Dependent Regret (Section 5)

C.1 Proof of Theorem 9

Fix $\ell \in [L]$ and let $P_\ell \doteq [s_\ell, e_\ell]$ denote the ℓ-th (stationary) phase. Fix also an arm a and suppose WLOG that arm a is the a-th arm to be evicted (i.e., at round t_a) in the sense of (4) of Definition 8. Let I_ℓ be the indices in $[a]$ such that $[t_{i-1}, t_i - 1]$ intersects phase P_ℓ. Next, we have the regret contribution to our formula (5) of arm a in phase P_ℓ is:

$$\sum_{i \in I_\ell} \sum_{t \in [t_{i-1}, t_i - 1] \cap P_\ell} \frac{\delta_\ell(a)}{|\mathcal{S}_t|} \cdot \mathbf{1}\{\delta_\ell(a) > 0\} = \sum_{i \in I_\ell} \frac{|[t_i, t_{i+1} - 1] \cap P_\ell| \cdot \delta_\ell(a)}{|\mathcal{S}_t|} \cdot \mathbf{1}\{\delta_\ell(a) > 0\}$$
$$\leq \delta_\ell(a) \cdot \mathbf{1}\{\delta_\ell(a) > 0\} \sum_{t \in P_\ell \cap [1, t_a - 1]} \frac{1}{|\mathcal{S}_t|}.$$

Next, we apply (4) of Definition 8 for $[s_1, s_2] = P_\ell \cap [1, t_a - 1]$:

$$\delta_\ell(a) \sum_{t \in P_\ell \cap [1, t_a - 1]} \frac{1}{|\mathcal{S}_t|} \leq C_2^2 \frac{\log(T) \sum_{t \in P_\ell \cap [1, t_a - 1]} |\mathcal{S}_t|^{-1}}{\delta_\ell(a) \sum_{t \in P_\ell \cap [1, t_a - 1]} |\mathcal{S}_t|^{-1}} = \frac{C_2^2 \log(T)}{\delta_\ell(a)}.$$

Plugging the above RHS bound into our earlier calculations, and then summing over arms a and phases ℓ gives us the desired regret bound.

C.2 Proof of Theorem 11

As noted in Definition 8, in a safe environment we may WLOG take $t_K \doteq T + 1$ assume there's a safe arm $a^\sharp \in \cap_{t=1}^T \mathcal{S}_t$. Let $\mathcal{S}_{T+1} \doteq \emptyset$. Then, we have:

$$\begin{aligned}
\sum_{i=1}^{K} \sum_{t=t_{i-1}}^{t_i - 1} \mathbb{E}_{a \sim \mathrm{Unif}\{\mathcal{S}_t\}}[\delta_t(a)] &= \sum_{i=1}^{K} \sum_{a \in [K]} \mathbf{1}\{a \in \mathcal{S}_{t_i - 1} \backslash \mathcal{S}_{t_i}\} \sum_{t=1}^{t_i - 1} \frac{\delta_t(a)}{|\mathcal{S}_t|} && (t_K = T + 1) \\
&\leq C_2 \sum_{i=1}^{K} \sum_{a \in [K]} \mathbf{1}\{a \in \mathcal{S}_{t_i - 1} \backslash \mathcal{S}_{t_i}\} \sqrt{\sum_{t=1}^{t_i - 1} \frac{\log(T)}{|\mathcal{S}_t|}} && \text{(from (4))} \\
&\leq C_2 \sqrt{K \sum_{i=1}^{K} \sum_{t=1}^{t_i - 1} \frac{\log(T)}{|\mathcal{S}_t|}} && \text{(Jensen's inequality)} \\
&\leq C_2 \sqrt{K \sum_{i=1}^{K} |\mathcal{S}_{t_{i-1}}| \sum_{t=t_{i-1}}^{t_i} \frac{\log(T)}{|\mathcal{S}_{t_{i-1}}|}} && \text{(from } \mathcal{S}_t \supseteq \mathcal{S}_{t+1}) \\
&\leq C_2 \sqrt{KT \log(T)}.
\end{aligned}$$

C.3 Proof of Theorem 12

First, similar to Proposition 3 of Suk and Kpotufe (2022), using Freedman's inequality, we establish a concentration error bound on our estimates $\hat{\delta}_t(a', a)$ (7).

Proposition 18. *Let $\mathcal{E}$ be the event that for all rounds $s_1 < s_2$ and all arms $a, a' \in [K]$:*

$$\left| \sum_{t=s_1}^{s_2} \hat{\delta}_t(a', a) - \mathbb{E}\left[\hat{\delta}_t(a', a) \mid \mathcal{H}_{t-1} \right] \right| \leq 10(e-1) \left(\sqrt{\sum_{s=s_1}^{s_2} \frac{\log(T)}{|\mathcal{A}_s|}} + \max_{s \in [s_1, s_2]} \frac{\log(T)}{|\mathcal{A}_s|} \right), \tag{13}$$

$$\left| \sum_{t=1}^{T} \delta_t(\pi_t) - \mathbb{E}[\delta_t(\pi_t) \mid \mathcal{H}_{t-1}] \right| \leq 10(e-1) \left(\sqrt{\log(T) \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \frac{\delta_t^2(a)}{|\mathcal{A}_t|}} + \log(T) \right). \tag{14}$$

where recall $\{\mathcal{H}_t\}_{t=1}^T$ is the filtration generated by $\{\pi_s, Y_s(\pi_s)\}_{s=1}^t$. Then, $\mathcal{E}$ occurs w.p. at least $1 - 1/T^2$.

Proof. Both (13) and (14) follow from Freedman's inequality (Beygelzimer et al., 2011, Theorem 1). ■

Now, note that since $\mathcal{A}_t$ is $\mathcal{H}_{t-1}$ measurable:

$$\forall a', a \in \mathcal{A}_t : \mathbb{E}\left[\hat{\delta}_t(a', a) \mid \mathcal{H}_{t-1}\right] = \frac{\delta_t(a', a)}{|\mathcal{A}_t|}$$
$$\mathbb{E}[\delta_t(\pi_t) \mid \mathcal{H}_{t-1}] = \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a)}{|\mathcal{A}_t|}.$$

Let a_i be the i-th arm to be evicted from the active set. Let $\hat{t}_0 \doteq 1$ and let $\hat{t}_1, \ldots, \hat{t}_K$ be the ordered eviction times of arms $a_1, \ldots, a_K$, or else let $\hat{t}_i = T + 1$ if arm a_i is never evicted. Using (14) and AM-GM inequality, we have:

$$\begin{aligned}
\sum_{t=1}^{T} \delta_t(\pi_t) &\leq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a)}{|\mathcal{A}_t|} + 10(e-1)\left(\sqrt{\log(T) \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \frac{\delta_t^2(a)}{|\mathcal{A}_t|}} + \log(T)\right) \\
&\leq c_2 \left(\log(T) + \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \frac{\delta_t(a)}{|\mathcal{A}_t|}\right) \\
&= c_2 \left(\log(T) + \sum_{i=1}^{K} \sum_{t=\hat{t}_{i-1}}^{\hat{t}_i - 1} \mathbb{E}_{a \sim \mathrm{Unif}\{\mathcal{A}_t\}}[\delta_t(a)]\right).
\end{aligned}$$

This gives the desired regret bound so long as we can argue that $\{\hat{t}_i\}_{i \in [K]}$ are valid eviction times.

First, we claim that, on event $\mathcal{E}$, the safe arm $a^\sharp$ cannot be evicted from $\mathcal{A}_t$. Suppose $a^\sharp$ is evicted from the active set using the eviction criterion (Line 5) over subinterval $[s_1, s_2]$. Then, since $\hat{\delta}_t(a', a) \in [-1, 1]$, we have:

$$\sum_{s=s_1}^{s_2} \frac{1}{|\mathcal{A}_s|} \geq \max_{a' \in [K]} \sum_{s=s_1}^{s_2} \frac{\hat{\delta}_s(a', a^\sharp)}{|\mathcal{A}_s|} \geq C_5 \sqrt{\sum_{s=s_1}^{s_2} \frac{\log(T)}{|\mathcal{A}_s|}} \implies \sum_{s=s_1}^{s_2} \frac{1}{|\mathcal{A}_s|} \geq C_5^2 \log(T).$$

In particular, this means for $C_5 > 1$ in Line 5: we will have

$$\sqrt{\sum_{s=s_1}^{s_2} \frac{\log(T)}{|\mathcal{A}_s|}} \geq \max_{s \in [s_1, s_2]} \frac{\log(T)}{|\mathcal{A}_s|}.$$

Combining the above with (13), we see that:

$$\sum_{s=s_1}^{s_2} \delta_s(a^\sharp) > c_3 \sqrt{\sum_{s=s_1}^{s_2} \frac{\log(T)}{|\mathcal{A}_s|}},$$

which violates the definition of the safe arm (Definition 10) for C_5 sufficiently large. Thus, we conclude that $a^\sharp \in \cap_{t=1}^{T} \mathcal{A}_t$.

We next decompose the (weighted) dynamic regret of any arm $a \in \mathcal{A}_{\hat{t}_i - 1}$ over subinterval $[s_1, s_2]$ via:

$$\sum_{s=s_1}^{s_2} \frac{\delta_s(a)}{|\mathcal{A}_s|} = \sum_{s=s_1}^{s_2} \frac{\delta_s(a^\sharp)}{|\mathcal{A}_s|} + \sum_{s=s_1}^{s_2} \frac{\delta_s(a^\sharp, a)}{|\mathcal{A}_s|}.$$

The first bound is order $\sqrt{\sum_{s=s_1}^{s_2} \frac{\log(T)}{|\mathcal{A}_s|}}$ via the definition of $a^\sharp$ and the second sum is also the same order by our concentration estimate (13) and eviction criterion (Line 5). Thus, $\{\hat{t}_i\}_{i \in [K]}$ are valid eviction times w.r.t. safe armsets $\{\mathcal{A}_t\}_{t=1}^{T}$ and for appropriately chosen C_2 in (4).

C.4 Proof of Theorem 13

We lower bound the regret iteratively by first designing the hard environment for $[t_0, t_1)$, then for $[t_1, t_2)$, and so on. The following theorem serves as a base template which we we can re-use on each period $[t_i, t_{i+1})$.

Theorem 19. *Fix a positive integer T, number of arms $K \in [2, T] \cap \mathbb{N}$ and a real number $R \in [0, \sqrt{TK}]$. Let $\mathscr{E}'$ be the class of all environments such that $\sum_{t=1}^{T} \mathbb{E}_{a\sim \mathtt{Unif}\{[K]\}}[\delta_t(a)] \leq R$, and such that T is a valid eviction time w.r.t. initial time 1 and threshold $C_2 = 1$ in (4). Then, for any algorithm π, we have:*

$$\sup_{\mathcal{E}\in\mathscr{E}'} R_{\mathcal{E}}(\pi, T) \geq \frac{1}{32e^{9/2}} \cdot R.$$

Proof. As in the proof of Theorem 3, we'll follow a Le Cam's method style of argument for showing minimax lower bounds in stationary bandits (e.g. Lattimore and Szepesvári, 2020, Theorem 15.2). We will refine the argument to show a more structured lower bound of order R over the class of problems with gap-dependent rate at most R.

Consider an environment $\mathcal{E}_0$ where $\mu_t(1) \doteq \frac{1}{2}$ and $\mu_t(a) \doteq \frac{1}{2} - \frac{R}{4T} \cdot \left(\frac{K}{K-1}\right)$ for all arms $a \neq 1$. Note that the bound $R \leq \sqrt{TK}$ and $T \geq K$ ensures $\mu_t(a) \in [0, 1]$.

One can verify this environment has the right gap-dependent rate and so belongs to the class $\mathcal{E}'$.

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim\mathtt{Unif}\{[K]\}}[\delta_t(a)] = \frac{R \cdot K}{4(K-1)} \leq R.$$

Now, by pigeonhole principle, there must exist an arm $a \neq 1$ for which the arm-pull count $N_T(a) \doteq \sum_{t=1}^{T} \mathbf{1}\{\pi_t = a\}$ satisfies $\mathbb{E}_{\mathcal{E}_0}[N_T(a)] \leq T/(K-1)$. Consider an alternative environment $\mathcal{E}_1$ whose mean rewards are identical to those of $\mathcal{E}_0$ except $\mu_t(a) \doteq \frac{1}{2} + \Delta$. For $\Delta \doteq \frac{R}{8T}$, this alternative environment also belongs to the class $\mathscr{E}'$ since for $K \geq 2$:

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim\mathtt{Unif}\{[K]\}}[\delta_t(a)] = \left(\frac{R \cdot K}{4(K-1)} + T \cdot \Delta\right) \cdot \left(\frac{K-1}{K}\right) + \frac{\Delta \cdot T}{K} < R.$$

Next, we observe the following regret lower bounds depending on whether the total arm-pull count $N_T(1)$ of arm 1 is larger than $T/2$:

$$\begin{aligned} R_{\mathcal{E}_0}(\pi, T) &\geq \frac{T}{2} \cdot \left(\frac{R \cdot K}{4 \cdot T \cdot (K-1)}\right) \cdot \mathbb{P}_{\mathcal{E}_0}(N_T(1) \leq T/2) \\ R_{\mathcal{E}_1}(\pi, T) &\geq \frac{T}{2} \cdot \Delta \cdot \mathbb{P}_{\mathcal{E}_1}(N_T(1) > T/2). \end{aligned}$$

By Bretagnolle-Huber inequality (Lemma 20), the above regret lower bounds give us:

$$\begin{aligned} R_{\mathcal{E}_0}(\pi, T) + R_{\mathcal{E}_1}(\pi, T) &\geq \frac{R}{16} \left(\mathbb{P}_{\mathcal{E}_0}(N_T(1) \leq T/2) + \mathbb{P}_{\mathcal{E}'}(N_T(1) > T/2)\right) \\ &\geq \frac{R}{32} \exp\left(-\mathrm{KL}(\mathcal{E}_0, \mathcal{E}_1)\right), \end{aligned}$$

where we use $\mathrm{KL}(\mathcal{E}_0, \mathcal{E}_1)$ to denote the KL divergence between the induced distributions on decisions and observations over T rounds in environments $\mathcal{E}_0$ and $\mathcal{E}_1$.

Next, we upper bound the KL between induced distributions which can be decomposed using chain rule (Lattimore and Szepesvári, 2020, Lemma 15.1):

$$\mathrm{KL}(\mathcal{E}_0, \mathcal{E}_1) = \mathbb{E}_{\mathcal{E}_0}[N_T(a)] \cdot \mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2} - \frac{R}{4T} \cdot \left(\frac{K}{K-1}\right)\right), \mathrm{Ber}\left(\frac{1}{2} + \Delta\right)\right).$$

By reverse Pinsker's inequality for Bernoulli random variables (Sason and Verdú, 2016, Remark 33),

$$\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2}-\frac{R}{4T}\cdot\left(\frac{K}{K-1}\right)\right),\mathrm{Ber}\left(\frac{1}{2}+\Delta\right)\right)\leq\frac{4}{\frac{1}{2}-\Delta}\cdot\left(\frac{R}{4T}\cdot\left(\frac{K}{K-1}\right)+\Delta\right)^2.$$

Now, since $\Delta=\frac{R}{8T}\leq\frac{1}{8}\sqrt{\frac{K}{T}}\leq\frac{1}{8}$, we have the above RHS is upper bounded by $\frac{25}{6}\cdot(R/T)^2$. Now, since R is at most $\sqrt{TK}$ and $\mathbb{E}_{\mathcal{E}_0}[N_T(a)]\leq T/(K-1)$, we have:

$$R_{\mathcal{E}_0}(\pi,T)+R_{\mathcal{E}_1}(\pi,T)\geq\frac{R}{32}\cdot\exp\left(-\frac{25T}{6(K-1)}\cdot\left(\frac{R}{T}\right)^2\right)\geq\frac{R}{32}\cdot e^{-25/12}$$

It's left to verify that round T is a valid eviction time for both environments $\mathcal{E}_0$ and $\mathcal{E}_1$, or that (4) holds for $\mathcal{S}_t\equiv[K]$. Indeed, for $\mathcal{E}_0$, we have

$$R\leq\sqrt{TK}\leq 2\sqrt{TK}\left(\frac{K-1}{K}\right)\implies\sum_{s=s_1}^{s_2}\sum_{a\in[K]}\frac{\delta_t(a)}{K}=\sum_{s=s_1}^{s_2}\left(\frac{R}{4T}\right)\cdot\left(\frac{K}{K-1}\right)\cdot\frac{1}{K}\leq\sqrt{\frac{s_2-s_1+1}{K}},$$

for any $[s_1,s_2]\subseteq[1,T]$. A similar calculation applies for environment $\mathcal{E}_1$. ■

Now, equipped with this base lower bound, to prove Theorem 13, we concatenate the above construction K times (each time removing an arm from the armset) to establish a lower bound of order R for any set of rounds $\{t_i\}_{i=1}^K$ by concatenating environments for K different eviction times.

First, note though that if $\{t_i\}_{i=1}^K$ are valid eviction times, then we must have by summing (4) over arms a and periods $[s_1,s_2]=[t_{i-1},t_i-1]$:

$$\sum_{i=1}^{K-1}\sum_{t=t_{i-1}}^{t_i-1}\mathbb{E}_{a\sim\mathrm{Unif}\{\mathcal{S}_t\}}[\delta_t(a)]\leq\sum_{i=1}^{K-1}\sqrt{(t_{i+1}-t_i)\cdot(K+1-i)}.$$

Next, we can find a partition of $R=\sum_{i=1}^{K-1}R_i$ such that $R_i\leq\sqrt{(t_i-t_{i-1})\cdot(K+1-i)}$ for all $i\in[K-1]$: specifically, let $R_1\doteq\min\{\sqrt{(t_1-1)\cdot K},R\}$ and recursively define $R_i\doteq\min\{\sqrt{(t_i-t_{i-1})\cdot(K+1-i)},R-\sum_{j=1}^{i-1}R_j\}$. By virtue of $R\leq\sum_{i=1}^{K-1}\sqrt{(t_i-t_{i-1})\cdot(K+1-i)}$, we claim there must exist an index i for which $R_i=R-\sum_{j=1}^{i-1}R_j$, then all subsequent $R_{i+1},\ldots,R_{K-1}$ are zero and $\sum_{i=1}^{K-1}R_i=R$. If such an index did not exist, then we'd have $\sum_{i=1}^{K-1}\sqrt{(t_i-t_{i-1})\cdot(K+1-i)}\leq R$ by considering index $i=K-1$, which is a contradiction.

Next, note that Theorem 19 can be applied over each period $[t_{i-1},t_i-1]$ with a different armset $\mathcal{S}_{t_i}$ of size $K+1-i$ to obtain a lower bound of order R_i. Letting $\mathcal{S}_1=[K]$, each armset $\mathcal{S}_{t_i}$ will be defined to randomly exclude an arm in $\mathcal{S}_{t_{i-1}}$.

We next claim that our concatenated environments lie in the class $\mathcal{E}$. First, note that:

$$\sum_{i=1}^{K-1}\sum_{t=t_{i-1}}^{t_i-1}\mathbb{E}_{a\sim\mathrm{Unif}\{\mathcal{S}_t\}}[\delta_t(a)]\leq\sum_{i=1}^{K-1}R_i=R.$$

Next, we claim that $\{t_i\}_{i=1}^K$ are valid eviction times. This follows in a similar fashion to the proof of Theorem 19. Consider a generic subinterval $[s_1,s_2]\subseteq[1,t_i-1]$ and break it up according to the periods $[t_{j-1},t_j]$ which intersect it. Then, to show (4), it remains to show for $[s_{1,j},s_{2,j}]\doteq[s_1,s_2]\cap[t_{j-1},t_j-1]$,

$$\sum_{j:[t_{j-1},t_j-1]\cap[s_1,s_2]\neq\emptyset}\left(\frac{R_j}{2\cdot(t_j-t_{j-1})}\right)\cdot\frac{(s_{2,j}-s_{1,j})}{K+1-j}\leq\sqrt{\frac{s_{2,j}-s_{1,j}}{K+1-j}}.$$

The above inequality will follow from the construction that $R_j\leq\sqrt{(t_{j+1}-t_j)\cdot(K+1-j)}$ for each index $j\in[K]$.

C.5 Proof of Theorem 14

We first show (i). Let $a \in \text{argmin}_{a\in[K]} f_a(1/T)$ be an optimal arm at round 1. Let $m \doteq \lfloor \beta \rfloor$. Then, by Taylor's Theorem with Lagrange remainder, we have for all $x \in [0,1]$, there exists $\xi \in [0,1]$ such that:

$$\begin{aligned} f_a(x) &= |f_a(x) - f_a(1/T)| \\ &\le \sum_{k=1}^{m-1} \frac{|f_a^{(k)}(x)|}{k!} \cdot |x - 1/T|^k + \frac{|f_a^{(m)}(\xi)|}{m!} \cdot |x - 1/T|^m \\ &\le \lambda_{\max} \sum_{k=1}^{m} \frac{|x-1/T|^k}{k!} \\ &\le (e^{x-1/T} - 1) \cdot \sqrt{\frac{K}{T}}. \end{aligned}$$

Now, this means arm a must be safe in the sense of (6) for suitable constant C_3 since the gap at any round cannot exceed $\sqrt{K/T}$. This shows (i).

Next, for (ii), we have that if there exists $n \in \{0, \ldots, m\}$ such that $\lambda_n \doteq \sup_{a\in[K]} \sup_{x\in[0,1]} |f_a^{(n)}(x)| > \sqrt{K/T}$, then we can consider the lower bound (Theorem 3) for the (n, λ_n)-Hölder class which will force regret:

$$\Omega(T^{\frac{n+1}{2n+1}} \cdot K^{\frac{n}{2n+1}} \cdot \lambda_n^{\frac{1}{2n+1}}) \ge \Omega(\sqrt{KT}),$$

where the last inequality follows from $\lambda_n > \sqrt{K/T}$.

D Auxilliary Lemmas

Lemma 20 (Bretagnolle-Huber Inequality; Theorem 14.2 of Lattimore and Szepesvári (2020))**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \ge \frac{1}{2} \exp\left(-\mathrm{KL}(P, Q)\right),$$

where $A^c = \Omega \backslash A$ is the complement of A.