Smooth Non-stationary Bandits

Su Jia [1] **Qian Xie** [1] **Nathan Kallus** [1] **Peter I. Frazier** [1]

Abstract

In many applications of online decision making, the environment is non-stationary and it is therefore crucial to use bandit algorithms that handle changes. Most existing approaches are designed to protect against non-smooth changes, constrained only by total variation or Lipschitzness over time, where they guarantee $\tilde{\Theta}(T^{2/3})$ regret. However, in practice environments are often changing **smoothly**, so such algorithms may incur higher-than-necessary regret in these settings and do not leverage information on the rate of change. We study a non-stationary two-armed bandits problem where we assume that an arm's mean reward is a β-Hölder function over (normalized) time, meaning it is $(\beta-1)$-times Lipschitz-continuously differentiable. We show the first separation between the smooth and non-smooth regimes by presenting a policy with $\tilde{O}(T^{3/5})$ regret for $\beta = 2$. We complement this result by an $\Omega(T^{(\beta+1)/(2\beta+1)})$ lower bound for any integer $\beta \geq 1$, which matches our upper bound for $\beta = 2$.

1. Introduction

As a fundamental variant of the MAB problem, non-stationary bandits provide a middleground between the stochastic bandits (Lai et al., 1985) and adversarial bandits (Auer et al., 2002). In the standard non-stationary model (Besbes et al., 2014), the mean reward function is adversarially chosen in advance, and rewards are realized stochastically.

The adversary is confined by the *total variation* budget V: the mean reward function $r_a(t)$ of every arm a is required to be Lipschitz and have total variation $\sum_{t=1}^{T} |r_a(t) - r_a(t+1)| \leq V$. The problem under this framework is well understood. In particular, an optimal regret bound $\tilde{O}\left(V^{1/3}T^{2/3}\right)$ is known (Besbes et al., 2014), even when V is unknown (Cheung et al., 2019).

This model may nonetheless be overly pessimistic for some applications as it allows the adversary to **instantaneously** shock the reward function's slope. However, in many applications, the underlying environment changes in a smooth manner, e.g., temperature, demand for a seasonal products, economic factors, just to name a few.

This motivates us to consider adversaries constrained to choose reward functions that are **smooth** in time. We model the level of smoothness by borrowing a standard concept from non-parametric statistics, called the Hölder class. As formally defined in Section 2, a function is β-*Hölder* if the first $(\beta-1)$ derivatives exist and are Lipschitz. In particular, for $\beta = 1$, our model recovers the model of Besbes et al. (2014) with $V = O(1)$, which has an optimal $\tilde{\Theta}(T^{2/3})$ regret. By setting $\beta > 1$, we constrain the adversary more than in past literature. This motivates the following question: *Can we break the $\tilde{O}(T^{2/3})$ regret bound (i.e., optimal bound for $\beta = 1$) under smooth non-stationarity (i.e., $\beta \geq 2$)?*

In this work, we provide an affirmative answer by showing an $\tilde{O}(T^{3/5})$ upper bound for $\beta = 2$. A natural idea would be to predict the derivative of the reward function and then make decisions based on the predicted trend. Surprisingly, our algorithm, which achieves the optimal regret, does **not** use any derivative information. Moreover, we show that this bound is in fact nearly optimal:[1] for any integer $\beta \geq 1$, we show every policy has worst-case regret $\Omega(T^{(\beta+1)/(2\beta+1)})$.

1.1. Related Work

Past multi-armed bandit literature has recognized the importance of non-stationarity by considering several aspects: contextual information (Luo et al., 2018; Russac et al., 2019); uncertainty in the number of changes (Auer et al., 2019; Chen et al., 2019); and Bayesian prior information (Trovo et al., 2020). However, these papers make no smoothness assumptions. (The "smoothly changing setting" in Trovo et al. (2020) assumes Lipschitz reward functions that may

[1]Cornell University, Ithaca, New York, USA. Correspondence to: Su Jia <sj693@cornell.edu>, Qian Xie <qx66@cornell.edu>, Nathan Kallus <kallus@cornell.edu>, Peter I. Frazier <pf98@cornell.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]Unless stated otherwise, "nearly optimal" means optimal up to logarithmic factors.

be non-differentiable, and is hence different from our work.)

There is another line of related work that considers reward function generated by a stochastic process. For example, when the reward function is drawn from a known reflected Brownian motion with variance σ^2, the optimal regret is known to be $\tilde{O}\left(k\sigma^2\right)$. Other evolution models include other Gaussian processes (Grünewälder et al., 2010) and discrete Markov chains (Zhou et al., 2021).

Beyond bandits, model smoothness and the Hölder class are often studied in non-parametric statistics (Györfi et al., 2002; Tsybakov, 2004). In the MAB literature, Hölder smoothness has been considered in contextual bandits; see e.g., (Hu et al., 2020; Gur et al., 2022).

1.2. Our Contributions

As our first and most important contribution, we present the first **separation** for non-stationary bandits between the smooth ($\beta \geq 2$) and non-smooth (i.e., classical, or $\beta = 1$) regimes: we develop a policy that we show achieves $\tilde{O}(T^{3/5})$ regret when $\beta = 2$. This policy and regret bound also apply to instances with $\beta > 2$, since a β-Hölder function with $\beta > 2$ is also 2-Hölder. This is asymptotically lower than the optimal $\tilde{\Theta}(T^{2/3})$ regret bound for the classical $\beta = 1$ setting (Besbes et al., 2014), and shows that smoothness can be exploited to reduce regret. On the technical level, our analysis relies on an *amortization* which, as a byproduct, can also be applied to give an alternate proof for the $\tilde{O}(T^{2/3})$ upper bound when $\beta = 1$.

As our second contribution, we provide an $\Omega(T^{(\beta+1)/(2\beta+1)})$ lower bound on regret that is valid for for every integer $\beta \geq 1$. This shows that our upper bound for $\beta = 2$ is tight up to logarithmic factors and that our proposed policy is also nearly optimal. If our lower bound is tight for general β, then as $\beta \to \infty$, it would suggest that it is possible to achieve $T^{1/2+O(1/\beta)}$ regret, which would almost match that of the classical stationary bandits. If true, this would show that smoothness can be an effective replacement for stationarity as a way to achieve low regret.

Our third contribution is introducing the notion of smoothness in non-stationary online models, by borrowing a standard concept, Hölder smoothness, from non-parametric statistics. Our model fills the gap between the adversarial and model-based non-stationary models, and may open up a new direction in online decision making.

2. Formulation

Consider the following non-stationary bandit model. For each arm $a \in \{0, 1\}$ and time $t = 1, \ldots, T$ up to horizon T, let Z_a^t be an independent random variable taking values on $[-1, 1]$. An instance is given by the distribution of these variables. The *reward function* is $r_a(t) := \mathbb{E}[Z_a^t]$. Its dependence on t makes this bandit model non-stationary. We consider two cases: in the *one-armed case* $r_0(t) = 0$ for all t, while in the *two-armed case* we make no such restriction. We discuss generalizing to more arms in Section 7.

In each round $t \in [T]$,[2] the learner selects an arm A_t whereupon they observe and receive a *reward* $Y_t := Z_{A_t}^t$. Based on the observed past rewards, they then select the next arm A_{t+1}, with the goal of maximizing the cumulative sum of rewards. This decision-making process is called a *policy*.

The problem would be trivial if $r_a(\cdot)$'s were known: In each round t the optimal policy chooses any $a_t^* \in \arg\max_{a\in\{0,1\}} r_a(t)$ and collects reward $r^*(t) = \max_{a\in\{0,1\}}\{r_a(t)\}$ in expectation.

When $r_a(\cdot)$'s are unknown and the policy needs to learn the reward functions on the fly. A standard metric for assessing a policy A is called *regret*, defined as the difference between the expected total rewards of A and the optimum given knowledge of the $r_a(t)$'s.

Definition 2.1 (Regret)**.** The *regret*[3] of a policy A under instance $r = \{r_a(t)\}$ is defined as

$$\mathrm{Reg}(A, r) = \mathbb{E}\left[\sum_{t=1}^{T}\left(r^*(t) - Z_{A_t}^t\right)\right].$$

For a family $\mathcal{F}$ of instances, the *worst-case regret* of A is $\max_{r\in\mathcal{F}} \mathrm{Reg}(A, r)$. The *minimax regret* is the minimum achievable worst-case regret among all policies.

To define smoothness for a bandit instance, which is a collection of functions on a **discrete** domain $[T]$, we first define smoothness for functions on a **continuous** domain.

Definition 2.2 (Hölder Class)**.** For integers $\beta \geq 1$ and $L > 0$, we say a function $f : [0, 1] \to \mathbb{R}$ is *β-Hölder* and write $f \in \Sigma(\beta, L)$ if (i) f is $(\beta - 1)$-order differentiable, and (ii) $f^{(\beta-1)}$ and f are both L-Lipschitz.

For example, consider the values $\beta = 1$ and $\beta = 2$, which are the most important for our analysis. One can easily verify that $f \in \Sigma(1, L)$ if and only if f is L-Lipschitz, and that $f \in \Sigma(2, L)$ if and only if f is differentiable and f', f are L-Lipschitz.

A bandit instance is β-Hölder if its reward function can be "embedded" into a β-Hölder function in the following sense.

Definition 2.3 (Smooth Non-stationary Bandit Instance)**.** A non-stationary bandit instance $r = \{r_a(t)\}$ is called *β-Hölder*, if for each arm a, there exists a function $\mu_a \in \Sigma(\beta, L)$ with domain $[0, 1]$ such that $r_a(t) = \mu_a\left(t/T\right)$ for

[2] For any integer T, we write $[T] = \{1, 2, \ldots, T\}$.

[3] Our notion of regret is sometimes called the *dynamic regret*, since the arm in the benchmark may change over time. Distinct from this, there is a substantial literature where the regret is against the best **fixed** arm, *e.g.*, in adversarial bandits.

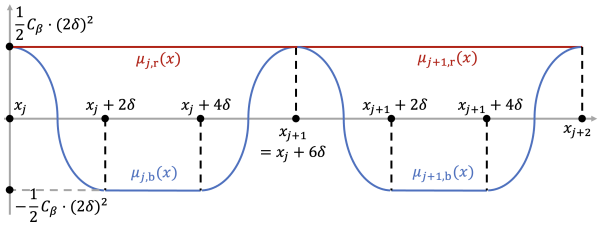

Figure 1. Illustration of the family $\mathcal{F}_\beta$ of reward functions. Here we show the "snapshots" of the curves on the two epochs $[x_j, x_{j+1}]$ and $[x_{j+1}, x_{j+2}]$. As can be seen from the figure, for any combination of red or blue curves, the change at any endpoint is **smooth** - both red and blue have 0 derivative at any x_j.

any $t \in [T]$. We denote by $\Sigma_k(\beta, L)$ the family of all β-Hölder instances with k arms.

For example, in the one-armed case, consider $r_1(t) = f(t/T)$ for $f(x) = |x - \frac{1}{2}|$. This instance is 1-Hölder but **not** 2-Hölder. In fact, for any differentiable function g with $r_1(t) = g(t/T)$, the derivative g' must change from -1 to $+1$ inside an interval of length $O(1/T)$. This means $|g''|$ is not bounded by any absolute constant.

We emphasize that in Definition 2.3, μ_a is defined on the *normalized* time scale $[0, 1]$ while r_a is defined on the $\{1, \ldots, T\}$. To avoid confusion, we will use x for the $[0, 1]$ scale and t for the $\{1, \ldots, T\}$ scale.

Finally, we note that (Manegueu et al., 2021) also considered *Hölder-continuous* rewards satisfying $|\mu_a(t) - \mu_a(t')| \leq |t - t'|^\beta$ for any $t, t' \in [0, 1]$ (their "α" is our "β"); see their page 11. This is only meaningful for $\beta \leq 1$, since otherwise the function is necessarily constant. To generalize to $\beta > 1$, our Hölder class requires **differentiability** and Hölder-continuity of the highest derivative.

In this sense, (Manegueu et al., 2021) considered $\beta < 1$, (Besbes et al., 2014) considered $\beta = 1$, and we are the first to consider $\beta > 1$, which is our main contribution. The leap from $\beta = 1$ to $\beta > 1$ is more challenging than from $\beta = 1$ to $\beta < 1$: the former must leverage derivatives, whereas the latter can still work with the functions themselves while better tracking general $\beta < 1$ exponents in arguments similar to the case of $\beta = 1$.

3. Lower Bounds

We show an $\Omega(T^{(\beta+1)/(2\beta+1)})$ lower bound on regret for the β-Hölder family of bandit instances. To show this bound, we construct a family $\mathcal{F}_\beta$ of β-Hölder instances and prove that every policy suffers the claimed regret on it.

We describe this family at a high level using Figure 1. Fix some $\delta > 0$ and partition $[0, 1]$ into *epochs* $[x_j, x_{j+1}]$ with $x_j = j \cdot 6\delta$ for $j = 0, 1, \ldots, m - 1$ where $m = \frac{1}{6\delta}$. For each reward function in this family, its restriction on each epoch is either constant, on the order of δ^β (colored red), or a *bowl-shaped* curve (colored blue). The family $\mathcal{F}_\beta$ consists of all such 2^m choices. A bowl curve b on $[x_j, x_{j+1}]$ is $(\beta - 1)$-times differentiable, with vanishing derivatives at x_j and x_{j+1}, which enables a **smooth** concatenation with a constant curve or another bowl curve.

As shown in Figure 1, over the two epochs $[x_j, x_{j+1}]$ and $[x_{j+1}, x_{j+2}]$, an instance can correspond to each of the following $2^2 = 4$ combinations of constant and bowl-shaped curves: a constant curve (all red); the curve constituted by two bowl-shaped curves, one on each epoch (all blue); and the two combinations of constant and bowl-shaped curves (one red then blue; the other blue then red). We formally define bowl-shaped curves and the family of instances $\mathcal{F}_\beta$ in 3.1, then use it to show our lower bound in 3.2.

3.1. Constructing the Family $\mathcal{F}_\beta$

Before constructing our family of instances $\mathcal{F}_\beta$ for the lower bound, the first question is: *do such bowl-shaped curves even* **exist***?* More precisely, is there a function which (i) has vanishing derivatives up to order $(\beta - 1)$ at the endpoints of an interval of length δ, and (ii) has maximum height $\Omega(\delta^\beta)$? We answer this affirmatively via an explicit construction.

Proposition 3.1. *Fix an integer $\beta \geq 1$. There exists a family $\{g_\varepsilon\}$ of $(\beta - 1)$-times continuously differentiable functions, with g_ε defined on $[0, \varepsilon]$, satisfying all of the following:*
(i) *(vanishing derivatives at the boundary)* $g_\varepsilon^{(j)}(0) = g_\varepsilon^{(j)}(\varepsilon) = 0$ *for any* $j = 1, \ldots, \beta - 1$,
(ii) *(monotonicity)* $g_\varepsilon' \geq 0$,
(iii) *(polynomial growth)* $g_\varepsilon(\varepsilon) = \Theta(\varepsilon^\beta)$ *as* $\varepsilon \to 0^+$. *Equivalently,* $g_\varepsilon(\varepsilon) = C_\beta(\varepsilon) \cdot \varepsilon^\beta$ *where* $C_\beta(\varepsilon) = \Theta(1)$,
(iv) *(Lipschitz derivatives)* $g_\varepsilon^{(\beta-1)}$ *is* 1*-Lipschitz.*

The proof of the above is rather technical and we defer the details to Appendix A.

As Figure 1 illustrates, a bowl-curve is obtained by connecting two rotated copies of g_δ with a constant function.

Definition 3.2 (Bowl-Shaped Curves)**.** Fix any integer $\beta \geq 1$ and let $\delta = \delta(\beta, T) = \left(2^{2(\beta+1)} C_\beta^2 T\right)^{-1/(2\beta+1)}$. For $j = 0, \ldots, m - 1$ and $x \in [0, 1]$, define

$$\mu_{j+1,\mathrm{r}}(x) = C_\beta(2\delta) \cdot (2\delta)^\beta \cdot \mathbb{1}_{[x_j, x_{j+1})}(x),$$

and $\mu_{j+1,\mathrm{b}}(x) =$

$$\begin{cases} -g_{2\delta}(x - x_j) + \frac{1}{2}C_\beta(2\delta)^\beta, & \text{if } x \in [x_j, x_j + 2\delta), \\ -\frac{1}{2}C_\beta(2\delta)^\beta, & \text{if } x \in [x_j + 2\delta, x_j + 4\delta), \\ g_{2\delta}(x - x_j - 2\delta) - \frac{1}{2}C_\beta(2\delta)^\beta, & \text{if } x \in [x_j + 4\delta, x_{j+1}), \\ 0, & \text{else.} \end{cases}$$

Each reward function in the family is specified by a binary vector that encodes its color in each epoch. The colors will be chosen adversarily by our lower bound in response to an algorithm's choices.

Definition 3.3 (The Family $\mathcal{F}_\beta$)**.** For any $v = (v_1, \ldots, v_m) = \{\mathrm{r}, \mathrm{b}\}^m$, let $\mu_v(x) = \sum_{j=1}^m \mu_{j,v_j}(x)$. We define $\mathcal{F}_\beta = \{\mu_v : v \in \{\mathrm{r}, \mathrm{b}\}^m\}$.

One can easily verify that for every $v \in \{\pm 1\}^m$ the function μ_v is β-Hölder. In particular, by property (i) in Proposition 3.1, if x is a multiple of δ, then for any $1 \le i \le \beta - 1$, the left and right order-i derivatives at x both become 0, ensuring a smooth transition.

The construction of bowl-shaped curves illustrates the role of β: for a fixed δ, the total variation of the bowl-shaped curve is only $O(\delta^\beta)$. In other words, the **more** smoothness we ask for, the **less** drastically the bowl curve can vary, reducing regret. This suggests that the lower bound and the optimal regret should decrease in β.

3.2. The Lower Bound

We are now ready to state the lower bound. Note our lower bound actually applies even if we restrict $Z_a^t \in \{\pm 1\}$.

Theorem 3.4 (Lower bound for Integer β)**.** *For any integer $\beta \ge 1$ and policy A, it holds that* $\mathrm{Reg}(A, v) > \frac{1}{24} \cdot 2^{-\beta} \left(C_\beta\right)^{-\frac{2\beta}{2\beta+1}} \cdot T^{\frac{\beta+1}{2\beta+1}}$.

As the key step, observe that for any $t \in [x_j + 2\delta, x_j + 4\delta]$, arm 0 is optimal under the blue curve and arm 1 is optimal under the red curve. We show that the probability of choosing the wrong arm in those rounds is at least $\frac{1}{2}$, under an adversarially chosen instance. This is true regardless of the algorithm and the shape of the reward function in the previous epochs.

Formally, consider two instances that are identical up to the $(j-1)^{\text{st}}$ epoch. For any prefix $u \in \{\mathrm{r}, \mathrm{b}\}^{j-1}$ and color $\chi \in \{\mathrm{r}, \mathrm{b}\}$, we will consider the instances $\mu_{u \oplus \mathrm{r}}$ and $\mu_{u \oplus \mathrm{b}}$ where $\oplus$ denotes vector concatenation. Write $t_j = x_j T$. We show the following in Appendix A.3.

Lemma 3.5 (Likely to Select a Wrong Arm)**.** *For any t in the j-th epoch $[t_{j-1}, t_j]$, any prefix $u \in \{\mathrm{r}, \mathrm{b}\}^{j-1}$ and any policy A, we have $\mathbb{P}_{u \oplus \mathrm{b}}[A_t = 1] + \mathbb{P}_{u \oplus \mathrm{r}}[A_t = 0] \ge \frac{1}{2}$.*

Observe that if the policy chooses a wrong arm at time $t \in [t_j + 2\delta T, t_j + 4\delta T]$, then an $\Omega(\delta^\beta)$ regret is incurred, leading to a high regret in this epoch.

We show that for every policy, there is an instance where $\Omega(\delta^\beta)$ regret is incurred in **every** epoch. Indeed, in $\mathcal{F}_\beta$, the shape of the reward function in previous epochs imposes no constraints on its shape in future epochs. Thus, given a policy, the adversary can choose the reward curve's shape in the next epoch (encoded red or bowl), whichever leads to higher regret. In other words, we have effectively m **separate** instances, each with time horizon $6\delta T$. We complete the proof by lower bounding the regret in each epoch separately using Lemma 3.5.

Proof of Theorem 3.4. Consider the regret on an epoch j under instance v,

$$\mathrm{Reg}_j(A, v) := \mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} \left(r_v^*(t) - Z_{A_t}^t\right)\right],$$

where $r_v^*(t) = \max\left\{0, \mu_v\left(\frac{t}{T}\right)\right\}$. Note that $\mathrm{Reg}_j(A, v)$ depends solely on the first j epochs, so we only need to specify the first j entries of v. Under this notation, by Lemma 3.5, for any prefix $u \in \{\mathrm{r}, \mathrm{b}\}^{j-1}$,

$$\begin{aligned}
&\mathrm{Reg}_j(A, u \oplus \mathrm{r}) + \mathrm{Reg}_j(A, u \oplus \mathrm{b}) \\
&\quad \ge \sum_{t=t_{j-1}+2\delta T}^{t_{j-1}+4\delta T} \left(\mathbb{P}_{u\oplus\mathrm{b}}[A_t = 1] + \mathbb{P}_{u\oplus\mathrm{r}}[A_t = 0]\right) \cdot \delta^\beta \\
&\quad \ge 2\delta T \cdot \frac{1}{2} \cdot \delta^\beta = \delta^{\beta+1} T.
\end{aligned}$$

Thus for some color $v_j \in \{\mathrm{r}, \mathrm{b}\}$, we have $\mathrm{Reg}_j(A, u \oplus v_j) \ge \frac{1}{2}\delta^{\beta+1}T$. Note that this inequality holds for any epoch j and prefix $u \in \{\mathrm{r}, \mathrm{b}\}^{j-1}$, so we can inductively construct a sequence $v \in \{\mathrm{r}, \mathrm{b}\}^m$ with $\mathrm{Reg}_j(A, v[j]) \ge \frac{1}{2}\delta^{\beta+1}T$ for each $j \in [m]$ where $v[j] = (v_1, \ldots, v_j)$. Summing over $j \in [m]$, we conclude that

$$\mathrm{Reg}(A, v) = \sum_{j=1}^m \mathrm{Reg}_j(A, v[j]) \ge m \cdot \frac{1}{2}\delta^{\beta+1}T = \frac{1}{12}\delta^\beta T.$$

Substituting $\delta = \left(2^{2(\beta+1)} C_\beta^2 T\right)^{-\frac{1}{2\beta+1}}$, we obtain

$$\mathrm{Reg}(A, v) > \frac{1}{24} \cdot 2^{-\beta} \left(C_\beta\right)^{-\frac{2\beta}{2\beta+1}} \cdot T^{\frac{\beta+1}{2\beta+1}}. \qquad \square$$

4. Algorithm and Analysis in One-Armed Case

We begin with the *one-armed* setting, i.e., $r_0 \equiv 0$, and consider a single arm with unknown reward function $r_1(t) = \mu_1(t/T)$ where $\mu_1 \in \Sigma(\beta, L)$. We can then suppress the subscripts and write $\mu(t) = \mu_1(t)$ and $r(t) = r_1(t)$. The generalization to the two-armed version is straightforward; see Section 5.

4.1. The Budgeted Exploration Algorithm

Consider the following *Budgeted Exploration (BE)* policy, formally defined in Algorithm 1. The policy is specified by two parameters, the *exploration budget* $B \ge 1$ and epoch size $\Delta \in (0, 1)$. The policy partitions the normalized timescale $[0, 1]$ into *epochs* $[x_i, x_{i+1}]$ where $x_i = i\Delta$ for each $i = 0, \ldots, \Delta^{-1} - 1$,[4] or equivalently, partitions

[4] For simplicity we assume Δ^{-1} is an integer. Apparently, this assumption is not essential.

$\{1, \ldots, T\}$ into epochs $[t_i, t_{i+1}]$ where $t_i = x_i T$. In each epoch, the algorithm pulls the changing arm (i.e., arm 1) from the start of the epoch until either (i) the epoch ends, or (ii) the budget B is run out, whereupon the algorithm selects the static arm (i.e., arm 0) in all remaining rounds in the epoch.

Algorithm 1 Budgeted Exploration Policy $\mathrm{BE}(B, \Delta)$

1: **for** $i = 1, \ldots, \Delta^{-1}$ **do**
2: Select arm 1 from round t_i until round $t_i + S_i$ with $S_i = \min\{\tilde{S}_i, \Delta T - 1\}$ where

$$\tilde{S}_i = \min\left\{ s : \sum_{t=t_i}^{t_i+s} Z_1^t \le -B \right\}.$$

3: Then select arm 0 from round $t_i + S_i + 1$ till t_{i+1}.
4: **end for**

We show that in the one-armed case, for suitable B and Δ, the BE algorithm achieves optimal regret bounds for $\beta = 1, 2$, both matching the lower bounds in Section 3.

Theorem 4.1 (Optimal Regret Bound, $\beta = 1$)**.** *For some $B = \tilde{O}(T^{1/3})$ and $\Delta = \tilde{O}(T^{-1/3})$, we have* $\mathrm{Reg}\left(\mathrm{BE}(B, \Delta), \Sigma_2\left(1, L\right)\right) = O\left(L^{1/3} T^{2/3} \log^{1/3} T\right)$.

Remarkably, the following upper bound provides the first separation between smooth ($\beta \ge 2$) and non-smooth ($\beta = 1$) bandits (Besbes et al., 2015).

Theorem 4.2 (Optimal Regret Bound, $\beta = 2$)**.** *For some $B = \tilde{O}(T^{2/5})$ and $\Delta = \tilde{O}(T^{-1/5})$, we have* $\mathrm{Reg}\left(\mathrm{BE}(B, \Delta), \Sigma_k\left(2, L\right)\right) = O\left(L^{1/5} T^{3/5} \log^{2/5} T\right)$.

Although these bounds are achieved by the same algorithm, their analyses involve different rationales. Specifically, for $\beta = 2$ the analysis must utilize smoothness to reduce the regret over what is achievable when $\beta = 1$. We will therefore prove them separately in Section 4.3 and 4.4.

4.2. Preliminaries

Before presenting the proofs of Theorems 4.1 and 4.2, we first state and prove tools used in both proofs. We will focus on bounding the regret conditional on the following *clean event*, which will be shown to occur with high probability. Loosely, this is the event that the rewards in all sufficiently large time intervals obey Hoeffding's inequality.

Definition 4.3 (Clean Event)**.** For any arm a and rounds $t, t' \in [T]$, consider the event

$$\mathcal{C}_a^{t,t'} = \left\{ \sum_{s=t}^{t'} \left(Z_{a,s} - \mu_a(s)\right) \le \sqrt{6 \log T \cdot (t' - t)} \right\}.$$

We define the *clean event* as $\mathcal{C} = \bigcap_{a,t,t'} \mathcal{C}_a^{t,t'}$ where the intersection is over all arms a and all t, t' with $t' - t \ge 2 \log T$.

We next show $\mathcal{C}$ occurs with high probability.

Lemma 4.4 (Clean Event Occurs w.h.p.)**.** *For any T, it holds that $\mathbb{P}[\overline{\mathcal{C}}] \le T^{-1}$.*

Proof. By Hoeffding's inequality (Vershynin, 2018, Theorem 2.2.6), for any $1 \le t \le t' \le T$ with $t' - t \ge 2 \log T$, taking $\delta = \sqrt{6 \log T \cdot (t' - t + 1)}$, we have

$$\begin{aligned}\mathbb{P}\left(\overline{\mathcal{C}_a^{t,t'}}\right) &\le \exp\left(-\frac{1}{2(t'-t+1)} \cdot 6 \log T \cdot (t' - t + 1)\right) \\ &= T^{-3}.\end{aligned}$$

There are at most T^2 combinations of t, t', so by the union bound, we have

$$\mathbb{P}[\overline{\mathcal{C}}] = \mathbb{P}\left[\bigcup_{a,t,t'} \overline{\mathcal{C}_a^{t,t'}}\right] \le \sum_{a,t,t'} \mathbb{P}\left(\overline{\mathcal{C}_a^{t,t'}}\right) \le T^{-1}. \quad \square$$

4.3. Proof of Theorem 4.1

Before delving deep into our focus, the $\beta = 2$ case, we analyze the $\beta = 1$ case as a warm-up. We will show a stronger statement than Theorem 4.1.

Proposition 4.5. *Suppose $6\Delta T \log T \le B^2$. Then*

$$\mathrm{Reg}\left(\mathrm{BE}(\Delta, B), \Sigma_1(1, L)\right) \le \Delta^{-1} \cdot (1 + L\Delta^2 T + B).$$

We then obtain Theorem 4.1 by selecting

$$\Delta = L^{-2/3} T^{-1/3} \log^{1/3} T \text{ and } B = L^{-1/3} T^{1/3} \log^{2/3} T.$$

At a high level, on each epoch the function μ either (i) is always positive, (ii) is always negative or (iii) has a unique crossing, i.e., $\mu(x) = 0$ for some x. We will bound the regret for these three types of epochs, referred to as *positive*, *negative* and *crossing* epochs, in Lemma 4.7, Lemma 4.8 and Lemma 4.9 respectively, and then combine them to complete the proof.

We first consider a positive epoch i. In this case, the optimal arm is arm 1, which coincides with the choice of the BE policy before the epoch's stopping rule is triggered. Thus, there is no regret in this epoch before time $t_i + S_i$. This can be formally shown by rephrasing Wald's classical identity as follows. Recall that (Z_1^t) are the rewards of arm 1.

Lemma 4.6 (Wald's Identity, Rephrased)**.** *For any epoch i, we have $\mathbb{E}\left[\sum_{t=t_i}^{t_i+S_i} Z_1^t\right] = \mathbb{E}\left[\sum_{t=t_i}^{t_i+S_i} r(t)\right]$.*

As a result, we only need to bound the regret incurred after $t_i + S_i$. We will do this by bounding the probability that the budget is ever run out, i.e., $S_i < \Delta T$, as detailed in the next lemma. Recall that $R_t = \max\{0, r(t)\} - Z^t_{A_t}$ is the regret in round t.

Lemma 4.7 (Regret on Positive Epochs)**.** *Suppose $\mu(x) > 0$ for $x \in [x_i, x_{i+1}]$. Then, whenever $B^2 \geq 6\Delta T \cdot \log T$, the regret of $\mathrm{BE}(B, \Delta)$ on epoch i satisfies $\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}} R_t\right] \leq 1$.*

The above essentially follows from the definition of the clean event. When $\mathcal{C}$ occurs, the cumulative reward up to the first s rounds in this epoch lies within an interval of width $w(s) \lesssim \sqrt{s}$ of the mean, which is positive.[5] Further, by the assumption that $B \gtrsim \sqrt{\Delta T}$ we have $w(s) \leq B$ whenever $s \leq \Delta T$. We defer the proof to Appendix B.2

Now we turn to the negative epochs.

Lemma 4.8 (Regret on Negative Epochs)**.** *If $\mu(t) < 0$ for $x \in [x_i, x_{i+1}]$, then the regret on epoch i satisfies $\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}} R_t\right] \leq B + 1$.*

We defer the details to Appendix B.3. The above follows from the definition of the stopping time S_i. In fact, if the process never stops until the end of epoch, then the cumulative reward is above $-B$. If it does stop at some $s < \Delta T$, then the cumulative reward is **just** below $-B$, and hence above $-(B+1)$ since $|Z^t_1| \leq 1$.

Finally we consider crossing epochs. The following essentially follows from the Lipschitzness of μ.

Lemma 4.9 (Regret on Crossing Epochs)**.** *Let j be a crossing epoch, i.e., $\mu(\tilde{x}) = 0$ for some $\tilde{x} \in [x_j, x_{j+1}]$. Then, the regret in this epoch satisfies $\mathbb{E}\left[\sum_{t=t_j}^{t=t_{j+1}} R_t\right] \leq 2L\Delta^2 T$.*

Proof. By Lipschitzness of μ, we have $|\mu(x)| = |\mu(x) - \mu(\tilde{x})| \leq L\Delta$ whenever $x_j \leq x \leq x_{j+1}$. In the original time scale, this means $|r(t)| \leq L\Delta$ whenever $t_j \leq t \leq t_{j+1}$, and hence

$$\sum_{t=t_j}^{t_{j+1}} |r(t)| \leq \Delta T \cdot L\Delta = L\Delta^2 T.$$

To connect the above with the regret, observe that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} R_t\right] &= \mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} (\max\{0, r(t)\} - Y_t)\right] \\ &\leq \sum_{t=t_j}^{t_{j+1}} |r(t)| - \mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} Z^t_{A_t}\right].\end{aligned}$$

Moreover, for each t we have $|\mathbb{E}[Z^t_{A_t}]| \leq |r(t)|$, so

$$\mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} R_t\right] \leq 2\sum_{t=t_j}^{t_{j+1}} |r(t)| \leq 2L\Delta^2 T.$$

□

We are now ready to show the $\tilde{O}(T^{2/3})$ upper bound. To suppress notations, we will subsequently write $R[i] = \mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}} R_t\right]$ as the regret on epoch i.

Proof of Proposition 4.5. Let $J_+, J_-, J_{\mathrm{x}} \subseteq \{1, \ldots, \Delta^{-1}\}$ be the subsets of positive, negative and crossing epochs. Note that $J_+ \cup J_- \cup J_{\mathrm{x}} = \{1, \ldots, \Delta^{-1}\}$, so the total regret can be decomposed as

$$\sum_{i=1}^{1/\Delta} R[i] = \sum_{i \in J_-} R[i] + \sum_{i \in J_+} R[i] + \sum_{i \in J_{\mathrm{x}}} R[i]. \quad (1)$$

By Lemma 4.7, Lemma 4.8 and Lemma 4.9, whenever $6\Delta T \log T \leq B^2$, we have

$$\begin{aligned}(1) &< |J_+| \cdot 1 + |J_-| \cdot (B+1) + |J_{\mathrm{x}}| \cdot 2L\Delta^2 T \\ &\leq \Delta^{-1} \cdot (1 + B + L\Delta^2 T).\end{aligned}$$

□

4.4. Proof of Theorem 4.2

We next show the $\tilde{O}(T^{3/5})$ regret for $\beta = 2$. We will show the following bound for generic policy parameters, which implies Theorem 4.2 by choosing

$$\Delta = L^{-2/5} T^{-1/5} \log^{1/5} T \text{ and } B = L^{-1/5} T^{2/5} \log^{3/5} T.$$

Proposition 4.10. *Suppose $6\Delta T \log T \leq B^2$, then*

$$\mathrm{Reg}\left(\mathrm{BE}(\Delta, B), \Sigma_1(2, L)\right) \leq 2\Delta^{-1} \cdot (L\Delta^3 T + B).$$

The proof strategy is similar to the $\beta = 1$ case. Specifically, note that Lemma 4.7 and Lemma 4.8 do **not** rely on the smoothness of μ, so they also hold for $\beta = 2$. However, as the key difference, we will derive a more fine-grained regret bound for the crossing epochs. To state this result, we classify the epochs based on whether μ' ever vanishes on them.

Definition 4.11 (Stationary Points and Stationary Epochs)**.** A point $s \in [0, 1]$ is said to be *stationary* if $\mu'(s) = 0$. An epoch is said to be *stationary* (or *non-stationary* otherwise) if it contains a stationary point.

As the key step, we need the following regret bound on the crossing epochs, which resembles Lemma 4.9 but is more refined as it crucially relies on the fact that $\beta = 2$.

[5] When illustrating high level ideas, we use $A \lesssim B$ to denote $A = \tilde{O}(B)$.

Lemma 4.12 (Key Lemma: Regret on Crossing Epochs)**.** *Let j be a crossing epoch and $j+\ell$ be a stationary epoch, with ℓ being possibly negative or 0. Moreover, suppose every epoch between them is non-stationary, i.e., epoch i is non-stationary whenever $(j+\ell-i)\cdot(j-i) < 0$. Then, the regret in epoch j satisfies $\mathbb{E}\left[\sum_{t=t_j}^{t=t_{j+1}} R_t\right] \le 2L\cdot(|\ell|+1)\cdot\Delta^3 T$.*

Proof. We assume $\ell \ge 0$ w.l.o.g. As the key observation, we first claim that $|\mu'| = O((\ell+1)\Delta)$ on epoch j. In fact, since epoch $(j-\ell)$ is stationary, there exists an $s \in [x_{j-\ell}, x_{j-\ell+1}] \subseteq [0,1]$ with $\mu'(s) = 0$. Moreover, since μ' is Lipschitz, for any $x \in [x_j, x_{j+1}]$ we have

$$|\mu'(x)| = |\mu'(x) - \mu'(s)| \le L\cdot|x-s| \le L\cdot(\ell+1)\Delta.$$

We next claim that $|r(t)| = O((\ell+1)\Delta^2)$ on epoch j. In fact, let $\tilde{x} \in [x_j, x_{j+1}]$ be any crossing, i.e., $\mu(\tilde{x}) = 0$. Then for any $x \in [x_j, x_{j+1}]$, by the mean value theorem, for some ζ between $\tilde{x}$ and x, i.e., $(\zeta - x)\cdot(\zeta - \tilde{x}) \le 0$, it holds that

$$|\mu(x)| = |\mu(x) - \mu(\tilde{x})| = |\mu'(\zeta)\cdot(\tilde{x} - x)|.$$

By the previous claim, i.e., $|\mu'| \le L\cdot(\ell+1)\Delta$ on epoch j, the above implies that

$$|\mu(x)| \le L\cdot(\ell+1)\Delta\cdot|\tilde{x} - x| \le L\cdot(\ell+1)\Delta^2.$$

Translating this to the original time scale, we have $|r(t)| \le L\cdot(\ell+1)\Delta^2$ for any $t \in [t_j, t_{j+1}]$, and thus the claim holds.

Finally, summing over t's, we obtain

$$\sum_{t=t_j}^{t_{j+1}} |r(t)| \le L(\ell+1)\Delta^2\cdot\Delta T. \tag{2}$$

We use this to bound the regret. Note that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} R_t\right] &= \mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} (\max\{0,\mu(t)\} - Y_t)\right] \\ &\le \sum_{t=t_j}^{t_{j+1}} \big|\mu(t)\big| - \mathbb{E}\left[\sum_{t=t_j}^{t_{j+1}} Y_t\right]. \end{aligned}\tag{3}$$

Observe that for each t we have $|\mathbb{E}Y_t| \le |r(t)|$, so the above is bounded by $2\sum_{t=t_j}^{t_{j+1}} \big|r(t)\big|$. Combining this with (2), we conclude that (3) $\le 2L\cdot(\ell+1)\cdot\Delta^3 T$. □

The above lemma suggests that for the adversary to generate a high regret on a crossing epoch, the nearest stationary point must be **proportionally** far away. This motivates us to consider an *amortization* scheme: bucket the epochs $1,\dots,\Delta^{-1}$ into contiguous blocks (called *cycles*) separated at stationary epochs, and then show that in each cycle, the regret on each epoch is low **on average**.

This proof strategy, however, assumes there is at least one stationary point. Thus, as the final building block, we need to handle the corner case where μ has no stationary point. We prove the following in Appendix B.

Lemma 4.13 (Corner Case)**.** *Suppose $\mu'(x) \ne 0$ for all $x \in [0,1]$. Then,* $\mathrm{Reg}\left(\mathrm{BE}(B,\Delta),\mu\right) \le L\Delta^2 T + (B+1)\Delta^{-1}$.

Now we are ready to prove the main upper bound.

Proof of Proposition 4.10. If there is no stationary point, i.e., $\mu'(t) \ne 0$ for all $t \in [0,1]$, then the desired bound follows immediately from Lemma 4.13.

Otherwise, index the stationary epochs as $s_1 < \cdots < s_n$. Crucially, observe that for any j, there is at most one crossing epoch between s_j and s_{j+1}, say i_{x}, if it does exist. Then by Lemma 4.12,

$$R[i_{\mathrm{x}}] \le 2L\cdot(|i_{\mathrm{x}} - s_j| + 1)\cdot\Delta^3 T \le 2L\cdot(|s_{j+1} - s_j|)\cdot\Delta^3 T.$$

Combining the above with Lemma 4.7 and Lemma 4.8, the total expected regret on epochs s_j through s_{j+1} satisfies

$$\sum_{s_j \le i < s_{j+1}} R[i] \le (s_{j+1} - s_i)\cdot(2L\Delta^3 T + B + 1).$$

The above clearly also holds when there is no crossing between s_j and s_{j+1}. In fact, in this case, the $2L\Delta^3 T$ term disappears. Summing over $j = 1,\dots,n$, we have

$$\begin{aligned}\textstyle\sum_{i=1}^{1/\Delta} R[i] &\le \textstyle\sum_{j=1}^{n}\sum_{s_j \le i < s_{j+1}} R[i] \\ &\le \textstyle\sum_{i=1}^{n}(s_{j+1} - s_i)\cdot(2L\Delta^3 T + B + 1) \\ &\le 2n\cdot(L\Delta^3 T + B),\end{aligned}$$

and the desired bound follows by noticing that $n \le \frac{1}{\Delta}$. □

5. Two-Armed Setting

Algorithm 2 adapts the BE policy to the two-armed case. In each epoch it alternates between the two arms until the difference in the cumulative rewards exceeds B. It then chooses the empirically better arm until the epoch ends.

Its analysis is similar to the one-armed case by considering the gap function $G(x) = \mu_0(x) - \mu_1(x)$ in place of $\mu_1(x)$. The only difference is that positive and negative epochs are replaced by *non-crossing epochs* via the following lemma.

Lemma 5.1 (Non-crossing Epoch)**.** *Suppose $G(x) \ne 0$ for any $x \in [x_j, x_{j+1}]$. Then the regret in this epoch satisfies $\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}} R_t\right] \le B + O(1)$.*

This is essentially obtained by merging the two lemmas for positive and negative epochs in the previous section. W.l.o.g. assume $\Delta > 0$. First we show that the regret by time $2S_i$ is no greater than $B+1$, using the definition of S_i and the boundedness of the rewards, as in the proof of Lemma 4.8.

Algorithm 2 BE(B, Δ) Policy, Two-Armed Case

1: **for** $i = 1, \ldots, \Delta^{-1}$ **do**
2: Select arm 0 in rounds $t_i + 2k$ and arm 1 in rounds $t_i + 2k + 1$ for $k = 0, 1, \ldots, S_i$ where $S_i = \min\left\{\tilde{S}_i, \frac{1}{2}(\Delta T - 1)\right\}$ with

$$\tilde{S}_i = \min\left\{s : \left|\sum_{k=0}^{s}\left(Z_0^{t_i+2k} - Z_1^{t_i+2k+1}\right)\right| > B\right\}.$$

3: Let $\hat{A}$ be the arm with higher cumulative rewards from rounds t_i to $t_i + 2S_i + 1$. Select $\hat{A}$ from round $t_i + 2(S_i + 1)$ till t_{i+1}.
4: **end for**

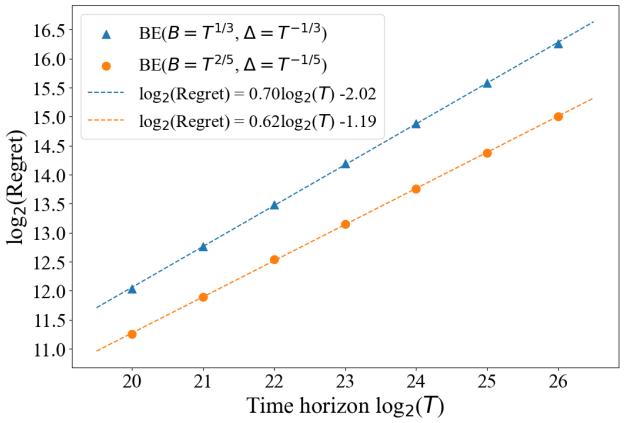

Figure 2. Log-log plot of the averaged regrets incurred by policy BE$(B = T^{1/3}, \Delta = T^{-1/3})$ and BE$(B = T^{2/5}, \Delta = T^{-1/5})$ as a function of the time horizon length T, with the linear relationship estimates. Here, the averaged regrets are calculated across randomly generated instances with sinusoidal mean rewards.

Then, we show that the expected regret after $2S_i$ is $O(1)$, as in the proof of Lemma 4.7. This can be done by considering the event that the better arm, i.e., arm 0, ever has cumulative rewards lower than arm 1 by B. Concentration bounds can show that this event occurs with low probability. Moreover, due to the boundedness of G, when this event occurs, the regret is also bounded, leading to $O(1)$ regret.

6. Experiments

We implemented our algorithm with simulations on synthetic data in the one-armed setting. We consider our BE policy where the parameters are chosen to be optimal for the non-smooth and smooth environments respectively. Formally, we consider the policy BE(B, Δ) where the tuple (B, Δ) is chosen to be $(T^{1/3}, T^{-1/3})$ for non-smooth and $(T^{2/5}, T^{-1/5})$ for smooth non-stationary environments.

We consider random *trigonometric* reward functions whose amplitudes, frequencies and phase shifts are randomly drawn. Specifically, in each instance, we have $r_0(t) = A$ and $r_1(t) = -A \cdot \sin(2\pi\nu t/T + \phi) + A$, where $\nu \sim \mathcal{U}_{[2.5,5]}$, $A \sim \mathcal{N}(0.25\nu^{-2}, 0.001)$ and $\phi \sim \mathcal{U}_{[0,2\pi]}$.

An astute reader may have noticed that A depends on ν when we generate the instances. This choice is actually quite natural. In fact, consider $\mu(x) = -A\sin(2\pi\nu x + \phi) + A$. Note that μ'' has a $\nu^2 A$ term, so by choosing A to scale as ν^{-2}, $|\mu''|$ becomes bounded by an absolute constant and hence $\mu \in \Sigma(2, L)$ for some $L = O(1)$.

We visualize the regret of the two policies via a log-log plot with time horizon $T = 2^j$ where $j = 20, 21, \ldots, 26$; see Figure 2. Theoretically, the **slope** of a log-log curve should equal the exponent of the cumulative regret. In fact, if the cumulative regret is cT^d, then the log-regret is $\log c + d \log T$. Our simulation shows that under smooth non-stationarity, the $T^{3/5}$-regret policy outperforms the $T^{2/3}$-regret policy. Moreover, the log-log curves have slope 0.70 and 0.62 respectively, which are close to their theoretical values.

7. Conclusions and Future Directions

In this paper, we presented *smoothly-varying* non-stationary bandits and demonstrated the first separation between the smooth and non-smooth case by showing we can break the $\tilde{O}(T^{2/3})$ regret lower bound for Lipschitz variation if we further assume Lipschitz-continuous differentiablity, attaining $\tilde{O}(T^{3/5})$ regret. We showed this upper bound is tight by establishing a lower bound of $\Omega(T^{(\beta+1)/(2\beta+1)})$ for any β-Hölder smooth variation.

One important direction is more than two arms. Algorithm 2 can be straightforwardly adapted to more arms: perform successive elimination in each epoch, until either the exploration budget is used up or the epoch ends. But it is not clear whether the regret has sublinear dependence on k.

We conjecture that the lower bounds can be matched for every integer $\beta \geq 3$ but this remains open. If this is true, it means that as smoothness increases we can obtain regret that approaches the optimal $\tilde{O}(\sqrt{T})$ regret of **stationary** bandits, since $\tilde{O}(T^{(\beta+1)/(2\beta+1)}) = T^{1/2+O(1/\beta)}$.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. 1846210. Peter Frazier was supported by AFOSR FA9550-19-1-0283 and FA9550-20-1-0351.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Auer, P., Gajane, P., and Ortner, R. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pp. 138–158. PMLR, 2019.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pp. 696–726. PMLR, 2019.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Grünewälder, S., Audibert, J.-Y., Opper, M., and Shawe-Taylor, J. Regret bounds for gaussian process bandit problems. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 273–280. JMLR Workshop and Conference Proceedings, 2010.

Gur, Y., Momeni, A., and Wager, S. Smoothness-adaptive contextual bandits. *Operations Research*, 70(6):3198–3216, 2022.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Hu, Y., Kallus, N., and Mao, X. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pp. 2007–2010. PMLR, 2020.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Luo, H., Wei, C.-Y., Agarwal, A., and Langford, J. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pp. 1739–1776. PMLR, 2018.

Manegueu, A. G., Carpentier, A., and Yu, Y. Generalized non-stationary bandits. *arXiv preprint arXiv:2102.00725*, 2021.

Russac, Y., Vernade, C., and Cappé, O. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Trovo, F., Paladino, S., Restelli, M., and Gatti, N. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

Tsybakov, A. B. Introduction to nonparametric estimation, 2009. *URL https://doi. org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Zhou, X., Xiong, Y., Chen, N., and Gao, X. Regime switching bandits. *Advances in Neural Information Processing Systems*, 34:4542–4554, 2021.

A. Omitted Proofs in Section 3

In this section we provide details for constructing the family in the lower bound proof.

A.1. Preliminaries: the Flock Transformation and the Pyramid

We need the notion of flocks to construct bowl-shaped curves. Pictorially, the flock transformation of a given function h (called the *base* function) is given by a sequence of copies of h side by side, each weighted by a constant. For example, the topmost subfigure in Figure 3 is a flock transformation of the pyramid-shaped base function.

Definition A.1 (Flock Transformation)**.** For any *base function* $h(x) : [0, w] \to \mathbb{R}$ and *weight* vector $v \in \mathbb{R}^\ell$, the v-*flock* is a function $F_v[h] : [0, \ell w] \to \mathbb{R}$ given by

$$F_v[h](x) = \sum_{i=1}^{\ell} v_i \cdot h\left((i-1)w + x\right).$$

We now specify the family of functions g_ε for constructing a bowl-shaped curve. We will set the highest derivative, i.e., $g_\varepsilon^{(\beta-1)}$, to be a flock transformation of the pyramid function whose weight vector is chosen from among the following *neutralizing* vectors.

Definition A.2 (Neutralizing Vectors)**.** Let $\nu^0 = 1$ and $\oplus$ be vector concatenation. For each integer $k \geq 1$, recursively define the k-th *neutralizing vector* as $\nu^k = \nu^{k-1} \oplus (-\nu^{k-1}) \in \{\pm 1\}^{2^k}$. A flock corresponding to a neutralizing vector is called a *neutral flock*.

As the name suggests, these vectors have the property that the sum of their entries on any dyadic interval is 0. Formally, $\sum_{j2^i+1}^{(j+1)2^i} \nu_i^k = 0$ for any $\nu = \nu^k$ and integers $i \geq 1$, $j \geq 0$. In our construction we set the highest derivative of g_ε to be a neutral flock of pyramid functions defined as follows.

Definition A.3 (Pyramid Function)**.** For any $w > 0$, define the w-*pyramid* function as

$$\Delta_w(x) = x \cdot \mathbb{1}_{\left[0, \frac{w}{2}\right]}(x) + \left(\frac{w}{2} - x\right) \cdot \mathbb{1}_{\left[\frac{w}{2}, w\right]}(x).$$

A.2. Proof of Proposition 3.1

We first describe the high level idea. We will prove Proposition 3.1 constructively by considering $g = g_\varepsilon$ obtained by iteratively integrating the pyramid function for $(\beta - 1)$ times. Figure 1 illustrates this idea for $\beta = 4$ using the pyramid function, with $\varepsilon = 4w$. Specifically, we start by setting the highest derivative, i.e., $g^{(3)}$, to be a neutral flock of pyramids; see the topmost subfigure. Then, as shown in the other two subfigures, the lower-order derivatives all **vanish** at the boundary points, i.e., at 0 and $4w$. Moreover, since $g^{(3)}$ is bounded by the function $f(x) = x$, we can verify that $g^{(i)}(\varepsilon) = O(\varepsilon^{4-i})$ for $i = 0, \ldots, 3$, in particular, we have $g(\varepsilon) = O(\varepsilon^4)$ as desired.

To formally define the family $\{g_\varepsilon\}$, we need the following.

Definition A.4 (Anti-derivative)**.** Consider any Lebesgue integrable function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$. We define $\Phi^0[f] = f$ and for any integer $\ell \geq 1$, the *level-ℓ anti-derivative* $\Phi^\ell[f] : \mathbb{R} \to \mathbb{R}$ is given by

$$\Phi^\ell[f](x) = \int_0^x \int_0^{x_\ell} \cdots \int_0^{x_2} f(x_1)\, dx_1 \ldots dx_\ell.$$

Now consider $g_\varepsilon = \Phi^{\beta-1}\left[F_{\nu^{\beta-2}}\left[\Delta_w\right]\right]$ with $w = w(\varepsilon) = 2^{-(\beta-1)}\varepsilon$. Note that the $(\beta-1)$-st derivative is $F_{\nu^{\beta-1}}[\Delta_w]$, which is 1-Lipschitz, so property (iv) holds trivially. Moreover, observe that $w2^{\beta-1} = \varepsilon$, and hence g_ε is supported on $[0, \varepsilon]$ as desired.

We next formally verify that $\{g_\varepsilon\}$ has the other properties claimed in Proposition 3.1. The next result, which connects the notions of anti-derivatives, flock transformation and neutralizing vectors, says that the anti-derivative of a neutral flock is still a neutral flock.

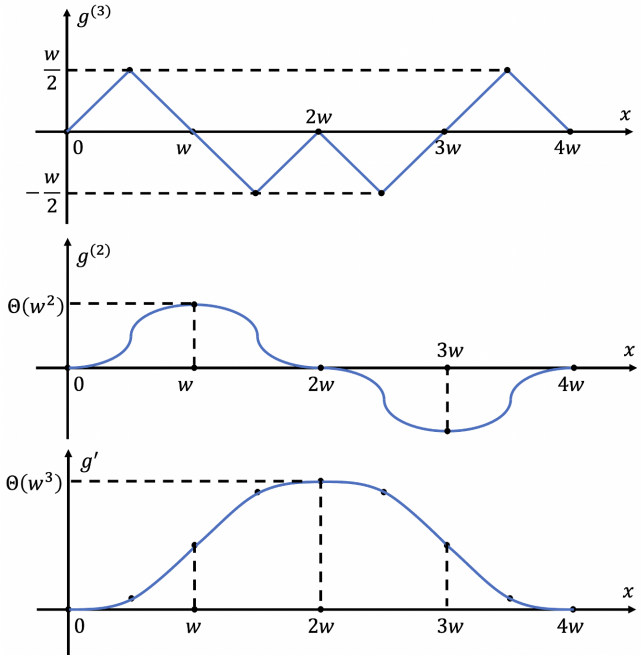

Figure 3. Illustration of g_ε for $\beta = 4$: $g^{(3)}$ is a pyramid flock. The function $g^{(2)}(x)$ is defined as the integration of $g^{(3)}$ from 0 to x. Similarly, $g^{(1)}(x)$ is the integration of g' from 0 to x.

Proposition A.5 (Anti-derivative Preserves Neutrality)**.** *Let $h_0 : [0, w] \to \mathbb{R}$ be any base function and let $j \geq 0$ be any integer. Then, for any $0 \leq \ell \leq k$, there exists some base function $h_\ell : [0, 2^\ell w] \to \mathbb{R}$ such that*

$$\Phi^\ell \left[F_{\nu^{\ell+j}} [h_0]\right] = F_{\nu^j}[h_\ell],$$

more precisely, $h_\ell = \Phi^\ell[F_{\nu^\ell}[h_0]]$.

To see the intuition, consider again $\beta = 4$ (readers can again refer to Figure 3) and consider pyramid base function $h_0 = \Delta_w$. In this case, the highest derivative is given by a ν^2-flock of pyramids, i.e.,

$$g^{(3)} = F_{\nu^2}[h_0]. \tag{4}$$

Now consider $g^{(2)}$. On one hand, as illustrated in the middle subfigure in Figure 3, $g^{(2)}$ is a ν^1-flock under the base function $h_1 = \Phi^1[F_{\nu^1}[\Delta_w]]$, i.e.,

$$g^{(2)} = F_{\nu^1}[h_1]. \tag{5}$$

On the other hand, by (4) we also have

$$g^{(2)} = \Phi^1[g^{(3)}] = \Phi^1[F_{\nu^2}[h_0]].$$

Combining with (5), we have

$$\Phi^1[F_{\nu^2}[h_0]] = F_{\nu^1}[h_1],$$

as claimed for $\ell = j = 1$.

Before formally showing this, we first state some basic properties of the flock transform. We use $\circ$ to denote the composition of mappings.

Lemma A.6 (Algebra of the Flock Transform)**.** *For any integers $i, j, k \geq 0$, it holds that*
(i) (Distributive Law) $(F_{\nu^i} \circ F_{\nu^j}) \circ F_{\nu^k} = F_{\nu^i} \circ (F_{\nu^j} \circ F_{\nu^k})$,
(ii) (Additive Law) $F_{\nu^i} \circ F_{\nu^j} = F_{\nu^{i+j}} = F_{\nu^j} \circ F_{\nu^i}$.

The proof of the above is rather straightforward and we leave it to the readers. We are now ready to prove Proposition A.5.

Proof of Proposition A.5. Consider induction on j. Wlog assume $w = 1$. The base case, $j = 0$, is trivially true, since F_{ν^0} is the identity mapping. Now consider $j \geq 1$. As the induction hypothesis, suppose the claim holds, i.e.,

$$\Phi^\ell \circ F_{\nu^{\ell+i}} = F_{\nu^i} \circ \Phi^\ell \circ F_{\nu^\ell},$$

for $i = 0, \ldots, j-1$. Denote by IH(i) the above identity. Then, for any base function h, it holds that

$$\begin{aligned}
\Phi^\ell \circ F_{\nu^{\ell+j}}[h] &= \Phi^\ell \circ F_{\nu^{\ell+j-1}}[F_{\nu^1}[h]] \\
&= F_{\nu^{j-1}} \circ \Phi^\ell F_{\nu^\ell}[F_{\nu^1}[h]] && \text{by IH}(j-1) \\
&= F_{\nu^{j-1}} \circ \Phi^\ell F_{\nu^{\ell+1}}[h] && \text{by Lemma A.6} \\
&= F_{\nu^{j-1}} \circ F_{\nu^1} \circ \Phi^\ell \circ F_{\nu^\ell}[h] && \text{by IH}(1) \\
&= F_{\nu^j} \circ \Phi^\ell \circ F_{\nu^\ell}[h], && \text{by Lemma A.6}
\end{aligned}$$

and the induction completes. □

We use Proposition A.5 to verify properties (i) and (ii) in Proposition 3.1. To verify that the derivatives do vanish at the endpoints, we need the following nice property of the neutralizing vectors.

Lemma A.7 (Symmetric Area Property)**.** *For any base function $h : [0, w] \to \mathbb{R}$ and integer $k \geq 1$, the ν^k-flock satisfies $\int_0^{2^k w} F_{\nu^k}[h](x)\, dx = 0$.*

Proof. Induction on k. For $k = 1$ this is obviously true. As the induction hypothesis, suppose this is true for some $k \geq 2$. Recall that by definition it holds that $\nu^k = \nu^{k-1} \oplus (-\nu^{k-1})$, so

$$\begin{aligned}
&\int_0^{2^{k+1}w} f_{w,\nu^{k+1}}(x)\, dx \\
&= \int_0^{2^k w} f_{w,\nu^k}(x)\, dx + \int_{2^k w}^{2^{k+1}w} f_{w,-\nu^k}(2^k w + x)\, dx \\
&= \int_0^{2^k w} f_{w,\nu^k}(x)\, dx - \int_0^{2^k w} f_{w,\nu^k}(x)\, dx \\
&= 0.
\end{aligned}$$

□

Proposition A.8 (Vanishing Derivatives at the Endpoints)**.** *Let $h : [0, w] \to \mathbb{R}$ be any base function and $H = F_{\nu^\ell}[h]$. Let $g = \Phi^\ell(H)$. Then, the function $g : [0, 2^\ell w] \to \mathbb{R}$ is ℓ-times continuously differentiable with $g^{(j)}(0) = g^{(j)}(2^\ell w) = 0$ for any $j = 1, \ldots, \ell$.*

Proof. By Proposition A.5, for any j, there exists base function $h_j : [0, 2^\ell w] \to \mathbb{R}$ such that $g^{(\ell-j)} = \Phi^j[H] = F_{\nu^{\ell-j}}[h_j]$. By Lemma A.7, it follows that

$$g^{(\ell-j-1)}(2^\ell w) = g^{(\ell-j-1)}(2^\ell w) - g^{(\ell-j-1)}(0) = \int_0^{2^\ell w} g^{(\ell-j)} = 0.$$

□

Proposition 3.1 then follows immediately since we have verified properties (i)-(iv).

A.3. Proof of Lemma 3.5

We first introduce some standard concepts and tools. For simplicity, for any instance $\mu : [0, 1] \to \mathbb{R}$, let $(Z_\mu^t)_{t \in [T]}$ be the reward vector under this instance.

Definition A.9 (KL-Divergence)**.** Let $X, Y \in \{\pm 1\}^n$ be random vectors, specified by probability mass functions $f_X, f_Y : \{\pm 1\}^n \to [0,1]$. The *Kullback-Leibler divergence* (or KL-divergence) is defined as

$$\mathrm{KL}(X,Y) = \sum_{v \in \{\pm 1\}^n} f_X(v) \log \frac{f_X(v)}{f_Y(v)}.$$

We show that at any time, the KL-divergence of the two random variables is on the order of the squared difference of their means.

Lemma A.10 (Bounding the KL-Divergence)**.** *For $i = 1, 2$, suppose random variable Z_i takes value on $\{\pm 1\}$ and has mean r_i. Then, when $|r_2| \le \frac{1}{2}$, we have $\mathrm{KL}(Z_1, Z_2) \le \frac{4}{3}(r_1 - r_2)^2$.*

Proof. By definition, we can write $Z_i = 2X_i - 1$ where $Z_i \sim \mathrm{Ber}\left(\frac{r_i+1}{2}\right)$ for $i = 1, 2$. Then, we have

$$\begin{aligned}
\mathrm{KL}(Z_1, Z_2) &= \mathrm{KL}(X_1, X_2) \\
&= \frac{r_1+1}{2} \ln \frac{r_1+1}{r_2+1} + \frac{1-r_1}{2} \ln \frac{1-r_1}{1-r_2} \\
&\le \frac{r_1+1}{2} \frac{r_1-r_2}{r_2+1} + \frac{1-r_1}{2} \frac{r_2-r_1}{1-r_2} \\
&\le \frac{(r_1-r_2)^2}{1-r_2^2} \\
&\le \frac{4}{3}(r_1-r_2)^2.
\end{aligned}$$

□

The following says that two instances with small KL-divergence are hard to distinguish between.

Theorem A.11 (Pinsker's Inequality)**.** *Let $X, Y \in \{\pm 1\}^n$ be two random vectors. For any event*[6] *E, we have $2(\mathbb{P}[Y \in E] - \mathbb{P}[X \in E])^2 \le \mathrm{KL}(X, Y)$.*

The chain rule characterizes the KL-divergence for random vectors, on which we will later apply Pinsker's inequality. The following can be found as Theorem 2.4 (b) in (Slivkins et al., 2019).

Theorem A.12 (Chain Rule for Product Distributions)**.** *Suppose $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \{\pm 1\}$ are independent. Consider $X = (X_i)$ and $Y = (Y_i)$. Then, $\mathrm{KL}(X, Y) = \sum_{t=1}^n \mathrm{KL}(X_t, Y_t)$.*

Proof of Lemma 3.5. For any color $\chi \in \{\mathrm{r}, \mathrm{b}\}$, denote by $Z_\chi = (Z_\chi^s)_{s=1}^{t-1}$ the reward vector under instance $u \oplus \chi$. Consider the event $E_t := \{A_t = 0\}$. By Pinsker's inequality (Theorem A.11) and the chain rule (Theorem A.12),

$$|\mathbb{P}[Z_\mathrm{r} \in E_t] - \mathbb{P}[Z_\mathrm{b} \in E_t]|^2 \le \frac{1}{2}\mathrm{KL}(Z_\mathrm{r}, Z_\mathrm{b}) = \frac{1}{2}\sum_{s=1}^{t} \mathrm{KL}(Z_\mathrm{r}^s, Z_\mathrm{b}^s) = 0 + \frac{1}{2}\sum_{s=t_{j-1}}^{t} \mathrm{KL}(Z_\mathrm{r}^s, Z_\mathrm{b}^s). \tag{6}$$

By the construction of $\mathcal{F}_\beta$ and Lemma A.10, for $t_{j-1} < s \le t$ we have $\mathrm{KL}(Z_\mathrm{r}^s, Z_\mathrm{b}^s) \le \frac{4}{3}\left(\frac{1}{2}C_\beta(2\delta)^\beta\right)^2$, and thus

$$(6) \le \frac{1}{2} \cdot \frac{4}{3}\left(\frac{1}{2}C_\beta(2\delta)^\beta\right)^2 \cdot (t - t_{j-1}) \le \frac{2^{2\beta}C_\beta^2}{3}\delta^{2\beta} \cdot 6\delta T, \tag{7}$$

where the last inequality follows since by definition, $t - t_{j-1} \le t_j - t_{j-1} = 6\delta T$. Finally, recall that $\delta = \left(2^{2(\beta+1)}C_\beta^2 T\right)^{-\frac{1}{2\beta+1}}$, so (7) gives $\left|\mathbb{P}[Z_\mathrm{r} \in E_t] - \mathbb{P}[Z_\mathrm{b} \in E_t]\right|^2 \le \frac{1}{4}$. Therefore,

$$\mathbb{P}[Z_\mathrm{r} \in E_t] + \mathbb{P}[Z_\mathrm{b} \in \overline{E_t}] = \mathbb{P}[Z_\mathrm{r} \in E_t] + 1 - \mathbb{P}[Z_\mathrm{b} \in E_t] \ge 1 - \left|\mathbb{P}[Z_\mathrm{r} \in E_t] - \mathbb{P}[Z_\mathrm{b} \in E_t]\right| \ge \frac{1}{2},$$

i.e., $\mathbb{P}_{u \oplus \mathrm{b}}[A_t = 1] + \mathbb{P}_{u \oplus \mathrm{r}}[A_t = 0] \ge \frac{1}{2}$. □

[6] In this work, by "event" we mean a Borel set.

B. Omitted Proofs in Section 4

B.1. Proof of Lemma 4.6

For simplicity fix i and write $S = S_i$. Observe that

$$\mathbb{E}\left[\sum_{t=t_i}^{t_i+S} Z_1^t\right] = \mathbb{E}\left[\sum_{t=t_i}^{t_i+S} \big(Z_1^t - r(t)\big)\right] + \mathbb{E}\left[\sum_{t=t_i}^{t_i+S} r(t)\right]. \tag{8}$$

Note that $Z_1^t - r(t)$'s are independent, mean 0, and S is a stopping time, so the partial sum $M_s := \sum_{t=1}^{s}(Z_1^t - r(t))$ is a martingale (w.r.t. the filtration induced by $\{Z_1^s\}$). By Wald's stopping time theorem, $\mathbb{E}[M_S] = 0$. Therefore, (8) becomes $0 + \mathbb{E}\left[\sum_{t=t_i}^{t_i+S} r(t)\right]$. □

B.2. Proof of Lemma 4.7

Write $S = S_i$. Since $\mu > 0$, the optimal policy always chooses arm 1 in this epoch. Recall that at time $t_i + S$ the BE algorithm switches to arm 0, the sub-optimal arm. We can thus simplify the regret as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}} (r(t) - Z_{A_t}^t)\right] \\
&= \mathbb{E}\left[\sum_{t=t_i}^{t_i+S} (r(t) - Z_1^t) + \sum_{t=t_i+S+1}^{t_{i+1}} (r(t) - Z_0^t)\right] \\
&= 0 + \mathbb{E}\left[\sum_{t=t_i+S+1}^{t_{i+1}} \big(r(t) - Z_0^t\big)\right] \\
&= \mathbb{E}\left[\sum_{t=t_i+S+1}^{t_{i+1}} r(t)\right],
\end{aligned} \tag{9}$$

where the last identity follows since Z_0^t's are mean 0 and independent of S. Further, since $t_{i+1} - t_i = \Delta T$ and $|r| \le 1$, we have

$$(9) \le \mathbb{P}[S = \Delta T - 1] \cdot 0 + \mathbb{P}[S < \Delta T - 1] \cdot \Delta T = \mathbb{P}[S < \Delta T - 1] \cdot \Delta T.$$

We conclude the proof by bounding $\mathbb{P}[S < \Delta T - 1]$. Consider the event $\{S = s\}$ where $s < \Delta T - 1$. We claim that the event $\{S = s\}$ would not occur conditional on the clean event $\mathcal{C}$. In fact, if $\{S = s\}$ occurs, we have $\sum_{t=1}^{s} Z_1^t < -B$. However, conditional on $\mathcal{C}$, we have

$$\Big|\sum_{t=t_i}^{t_i+s} (Z_1^t - r(t))\Big| \le \sqrt{6s \cdot \log T} < \sqrt{6\Delta T \log T},$$

and more explicitly,

$$\sum_{t=t_i}^{t_i+s} Z_1^t > \sum_{t=t_i}^{t_i+s} r(t) - \sqrt{6\Delta T \log T} \ge 0 - B,$$

where the last inequality follows since $6\Delta T \log T \le B^2$. It follows that $\mathbb{P}[\{S = s\} \cap \mathcal{C}] = 0$ for any $s < \Delta T - 1$, and hence

$$\mathbb{P}[S < \Delta T] = \mathbb{P}\left[\{S < \Delta T\} \cap \overline{\mathcal{C}}\right] \le \mathbb{P}\left[\overline{\mathcal{C}}\right] \le T^{-1}.$$

Therefore, $(9) \le \mathbb{P}[S < \Delta T] \cdot \Delta T \le T^{-1} \cdot \Delta T \le 1$. □

B.3. Proof of Lemma 4.8

Write $S = S_i$. In this case the optimal arm is arm 0, so we can simplify the regret as

$$\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}}(-Z^t_{A_t})\right] = -\mathbb{E}\left[\sum_{t=t_i}^{t_i+S} Z^t_1\right] - \mathbb{E}\left[\sum_{t=t_i+S+1}^{t_{i+1}} Z^t_0\right]. \tag{10}$$

Note that Z^t_0 is independent of S, so the second expectation is 0.

To analyze the first term, for any $s \geq 0$ define $X_s := -\sum_{t=t_i}^{t_i+s} Z^t_1$. Then by definition of S, on the event $\{S = s\}$ we have $X_{s-1} < B$. Since we assumed each the reward distribution to be $\{\pm 1\}$-valued, this implies that $X_s < B + 1$. Therefore,

$$-\mathbb{E}\left[\sum_{t=t_i}^{t_i+S} Z^t_1\right] = \mathbb{E}[X_S] = \sum_{s=1}^{\infty} \mathbb{E}\left[X_s \cdot \mathbb{1}(X = s)\right] < \sum_{s=1}^{\infty} \mathbb{P}[S = s] \cdot (B + 1) = (B + 1),$$

where the first identity follows from Lebesgue's Dominated Convergence Theorem and the boundedness of X_s for any fixed s. The claimed bound immediately follows by combining the above with (10). □

B.4. Proof of Lemma 4.13

Note that in this case μ has at most one crossing on its domain $[0, 1]$. As the trivial case, suppose there is no crossing, then the upper bound follows immediately by applying Lemma 4.7 or Lemma 4.8 on each epoch.

Now suppose there is exactly one crossing, say $\tilde{x} \in [x_{i_0}, x_{i_0+1}]$ for some integer i_0. Then, for any $x \in [x_{i_0}, x_{i_0+1}]$, by Lipschitzness, we have

$$|\mu(x)| = |\mu(x) - \mu(\tilde{x})| = |\mu'(\zeta) \cdot (x - \tilde{x})| \leq L\Delta.$$

Translating this to the original time scale, we have $|r(t)| \leq L\Delta$ whenever $t_{i_0} \leq t \leq t_{i_0+1}$. Therefore,

$$\sum_{t=t_{i_0}}^{t_{i_0+1}} |r(t)| \leq L\Delta \cdot (t_{i_0+1} - t_{i_0}) = L\Delta \cdot \Delta T. \tag{11}$$

Meanwhile, since any epoch $i \neq i_0$ is either negative or positive, by Lemma 4.7 and Lemma 4.8 we have $R[i] \leq B + 1$. Combining this with (11), the total regret is then bounded as

$$\sum_{i=1}^{1/\Delta} R[i] = R[i_0] + \sum_{i \neq i_0} R[i] < L\Delta^2 T + \Delta^{-1} \cdot (B + 1).$$

□