Abstract

We study a non-parametric multi-armed bandit problem with stochastic covariates, where a key driver of complexity is the smoothness with which the payoff functions vary with covariates. Previous studies have derived minimax-optimal algorithms in cases where it is a priori known how smooth the payoff functions are. In practice, however, advance information about the smoothness of payoff functions is typically not available, and misspecification of smoothness may severely deteriorate the performance of existing methods. In this work, we consider a framework where the smoothness is not known a priori, and study when and how algorithms may adapt to unknown smoothness. First, we establish that, in general, designing bandit algorithms that adapt to the unknown smoothness of payoff functions is impossible. We overcome this impossibility result by leveraging the notion of self-similarity, a concept from the statistics literature that is traditionally invoked to enable adaptive confidence intervals. Under a self-similarity assumption, we develop a policy for inferring the smoothness of the payoff functions using observations that are collected throughout the decision-making process, and establish that this policy matches (up to a logarithmic scale) the regret rate that is achievable when smoothness is known a priori. Finally, we extend our method to account for local notions of smoothness and show that, under reasonable assumptions, our method achieves performance characterized by the local complexity of the problem as opposed to its global complexity.

Keywords: Contextual multi-armed bandits, Hölder smoothness, self-similarity, non-parametric confidence intervals, non-parametric estimation

1 Introduction

Multi-armed bandits (MAB) (Thompson 1933, Robbins 1952) is a well-studied dynamic optimization framework that captures the trade-off between new information acquisition (exploration), and optimizing payoffs based on available information (exploitation). In many cases of interest, the decision maker also has access to covariate information that can be informative about the effectiveness of different actions (Woodroofe 1979). This setting is often referred to the contextual MAB problem; other authors have also referred to it as the bandit problem with side observations (Wang et al. 2005b), or the associative
bandit problem (Strehl et al. 2006). Following Woodroofe (1979), there has been a considerable amount of work on contextual bandits under parametric assumptions on the relationship between contexts and mean outcomes (parametric payoff functions); see Goldenshluger and Zeevi (2013) and Bastani and Bayati (2019) for some recent examples. In many applications, however, there is interest in studying non-parametric contextual MAB formulations, which make less structural assumptions, are typically more robust, and can be applied to a more general class of problems, especially, when less is known about the structure of payoff functions.

Inherently, the exploration-exploitation trade-off in a contextual setting is different compared to the classical MAB setting. This is mainly due to the fact that in a contextual setting, one may extrapolate collected information over the context space, which might significantly affect the amount of exploration required to guarantee “good” performance. The challenge of balancing information acquisition and instantaneous payoffs for personalized decisions in the presence of contextual information is fundamental in many application domains, including pricing (e.g., Cohen et al. 2016, Qiang and Bayati 2016, Ban and Keskin 2019, Bastani et al. 2019, Javanmard and Nazerzadeh 2019), product recommendations (e.g., Chu et al. 2011, Kallus and Udell 2016, Bastani et al. 2018, Agrawal et al. 2019, Gur and Momeni 2019), and healthcare (e.g., Tewari and Murphy 2017, Chick et al. 2018, Bastani and Bayati 2019, Zhou et al. 2019), to mention but a few.

In general, in contextual MAB, the smoothness of the payoff functions (describing how the payoff from selecting a certain action changes as a function of the context) determines the extent to which extrapolation can be done by the agent, which in turn influences the amount of exploration required to achieve a good performance. Therefore, this smoothness plays a key role in determining effective policy design as well as the best achievable performance. To demonstrate this principal in one concrete example, consider the problem of artwork selection on Netflix. A simple version of this problem is described in Chandrashekar et al. (2017) where two different artworks are available for recommending the movie Good Will Hunting (see the top parts of Figure 1). When Netflix recommends a title, it also needs to select an image to display along with the recommendation. Different images may generate different probabilities of playing the movie, and given the personal viewing history of the user, Netflix aims to select the imagery that maximizes the probability of playing the recommended title.

The bottom plots of Figure 1 illustrate two different scenarios of how the probability to play the title changes as a function of a single context, which equals the normalized difference of the viewer’s romance and comedy scores\(^1\) when Netflix uses image A or image B. For simplicity, this illustration assumes that

\(^1\)Romance and comedy scores are computed based on the viewing history of the users; see Chandrashekar et al. (2017) for more details
Figure 1: Top: Example of artwork selection on Netflix for recommending the movie Good Will Hunting (for details and discussion see Chandrashekar et al. 2017); Bottom: The probability of users to play the recommended title as a function of the normalized difference of their romance and comedy score (context) when either image A (dashed line) or image B (dotted line) is shown, in two different scenarios: (Bottom Left) users’ behavior changes linearly as a function of the context; Bottom Right: users’ behavior changes more abruptly as a function of the context. In each case $x^*$ denotes the context at which the optimal imagery switches.

in both cases the probability of playing is a monotone function of context. Let $x^*$ denote the context at which the optimal imagery switches. Particularly, if contexts fall in the interval $[0, x^*]$ image B is be the optimal imagery to display; otherwise, the optimal selection is image A. In the scenario illustrated on the bottom-left part of Figure 1 users’ behavior is smooth and changes linearly with respect to the context. In this case, any observation of users’ behavior, even if the user’s context is very close to 1 or 0, is informative and can be utilized for estimating the probability lines and the crossing point $x^*$. In contrast, the bottom-right part of Figure 1 depicts a scenario where the probability lines change more abruptly. In this case, observations whose contexts are very close to 1 or 0 are less informative and cannot be easily utilized for estimating the crossing point $x^*$. As a result, the second scenario requires more extensive experimentation over the context space in order to determine optimal decision regions.

Previous works on non-parametric contextual bandits typically assume prior knowledge of the worst-case smoothness of payoff functions. A standard approach is to assume that mean reward functions are $(\beta, L)$-Hölder (see Definition 2.1) for some known $\beta$ and $L$, and develop policies accordingly. Rigollet and Zeevi (2010) and Perchet and Rigollet (2013) develop minimax rate-optimal algorithms in the “rough” case where $0 < \beta \leq 1$ (in this case, the payoff functions are only assumed to be Lipschitz or rougher); more recently, Hu, Kallus, and Mao (2019) extended these results to the “smooth” case with $\beta > 1$. 

3
In practice, however, the class of functions to which payoff functions belong is often unknown, and misspecification of the smoothness parameters may cause significant deterioration in the performance of existing methods. On one hand, underestimating the smoothness of the payoff functions leads to excessive and unnecessary experimentation and therefore hurts performance. On the other hand, when the smoothness is overestimated, then the provided guarantees cease to hold and one might incur a linear regret. Therefore, there is an increasing interest in designing adaptive policies that achieve the best possible performance without any prior knowledge of the smoothness parameters. Addressing this challenge is the main focus of this paper.

**Main contributions.** Our contributions are to (1) formulate a non-parametric contextual MAB problem where the smoothness of payoff functions is a priori unknown; (2) analyze the complexity of adapting to smoothness, and show that smoothness-adaptivity is in general impossible; (3) identify self-similarity as a condition under which adaptivity is possible and design a policy that exploits this condition to adaptively guarantee rate-optimality; and (4) provide a framework for exploiting local smoothness in contextual MAB problems. More specifically, our contribution is along the following lines.

(1) **Modeling.** We formulate a non-parametric contextual MAB problem where the smoothness of payoff functions is a priori unknown. More precisely, the payoff functions will be assumed to belong to the Hölder class of functions with some unknown Hölder exponent $\beta \in [\underline{\beta}, \bar{\beta}]$. Our formulation allows for any arbitrary range of this smoothness parameter, and captures a large variety of real-world phenomena, yet maintains mathematical tractability.

(2) **Impossibility of adaptation.** We establish a lower bound on the best achievable performance when two different classes of payoff functions with different smoothness are considered simultaneously. Through this lower bound, we show that, in general, adaptively achieving rate-optimal performance uniformly over different classes of smooth payoff functions is impossible, despite the fact that smoothness-adaptive estimation of non-parametric functions is possible; see, e.g., Lepski (1992). Thus, this impossibility result highlights a fundamental difference between the complexities of non-parametric function estimation and non-parametric contextual MAB.

(3) **Policy design.** Despite the general impossibility of adapting to unknown smoothness, we show that when one assumes self-similarity of the payoff functions, then adaptive (nearly) rate-optimal performance becomes possible. Self-similarity is a well-known notion from statistics, and has been often invoked to enable construction of smoothness-adaptive confidence bands. In particular, constructing smoothness-adaptive confidence bands is impossible in general (Low 1997), but becomes possible under the self-similarity assumption (Giné and Nickl 2010). Here, we propose a policy for inferring the
worst-case global smoothness of the problem at hand and show that, given self-similarity, this policy lets us effectively estimate the smoothness of payoff functions. Furthermore, we show that the regret incurred by our policy matches (up to a logarithmic factor) the minimax regret that could be achieved if smoothness parameters were known a priori.

At a high level, our analysis implies that the complexity of the smoothness-adaptive non-parametric contextual MAB problem is more closely related to the problem of constructing smoothness-adaptive non-parametric confidence bands, relative to the complexity of smoothness-adaptive non-parametric point estimation. This highlights that constructing efficient contextual MAB policies essentially involves making accurate predictions. Notably, popular MAB policies (UCB, Thompson sampling, etc.) explicitly or implicitly rely on accurate confidence intervals or other forms of uncertainty quantification to guide exploration. Our results demonstrate that the complexity of adapting to smoothness in the contextual MAB problem is primarily driven by this required uncertainty quantification step.

(4) **Local complexity.** We advance beyond adapting to global smoothness of payoff functions and design a policy that accounts for the local complexity of the problem by adapting to how smooth payoff functions behave *locally* when the notion of self-similarity is leveraged in every local region. This policy is based on an incremental refinement of the covariate space, and relies on inferring the local smoothness of payoff functions in order to regulate its exploration rate and refinement policy.

### 1.1 Related literature

#### Parametric and non-parametric approaches in contextual multi-armed bandits.

Most of the literature on contextual MAB assumes parametric payoff functions. These studies can be categorized based on their assumption regarding the way contexts are realized. Some researchers have studied this setting when contexts are drawn independently form the same distribution. For example, [Goldenshluger and Zeevi (2013), Bastani et al. (2017), and Bastani and Bayati (2019)] consider linear payoff functions. On the other hand, [Langford and Zhang (2008) and Dudik et al. (2011)] study the problem of finding the best mapping from contexts to arms among a finite set of hypotheses. In addition, [Wang et al. (2005b)] considers a general relationships between the parameters of payoff functions and contexts; their results were later extended by [Wang et al. (2005a)] to the case where contexts are realized according to stochastic non-i.i.d. processes.

In contrast, some studies consider the case that contexts are chosen by an *oblivious adversary* whose actions are independent of the agent’s past actions. For example, [Abe and Long (1999), Auer (2002), and Chu et al. (2011)] consider linear payoff functions, and [Filippi et al. (2010)] consider generalized linear
relationships between contexts and rewards. Furthermore, Agrawal and Goyal (2013) extend the study of contextual MAB with linear payoffs beyond oblivious adversaries and, using a posterior sampling method, addresses the case where an adaptive adversary, who takes actions based on the agent’s past actions. Finally, there is also a literature stream known as MAB with expert advice, where contexts are realized in the form of advice from a pool of experts (Auer et al. 2002, Maillard and Munos 2011).

Aside from these parametric approaches, some researchers have addressed this problem from a non-parametric point of view to account for general relationships between covariates and mean rewards. Yang and Zhu (2002) is the first work to initiate this line of research. They combined an $\epsilon$-greedy-type policy with non-parametric estimation methods such as nearest neighbors to achieve strong consistency, which ensures the total reward collected by the agent is almost surely asymptotically equivalent to those obtained by always pulling the best arm. Following this work, some researchers have proved stronger results in this setting regarding regret rate. Rigollet and Zeevi (2010) introduced the UCBogram policy which decomposes the covariates space into bins and follows a traditional UCB policy in each bin separately. Perchet and Rigollet (2013) improved upon this result by introducing the Adaptively Binned Successive Elimination policy, which implements an increasing refinement of the covariate space and achieves the minimax regret rate. Recently, Hu et al. (2019) extended this framework to the case of smooth differentiable functions. Finally, Reeve et al. (2018) proposes the kNN-UCB policy that achieves the minimax regret rate and also adapts to the intrinsic dimension of data. All these studies, however, assume that the smoothness of the payoff functions is known a priori.

**Adaptive non-parametric methods.** For the general theory on adaptive non-parametric estimation, we refer the readers to Lepskii (1992). This line of research is rich and includes various approaches. For example, Donoho and Johnstone (1994), Donoho et al. (1995), and Juditsky (1997) deploy techniques based one wavelets, Lepski et al. (1997) proposes a kernel-based method, and Goldenshluger and Nemirovski (1997) develops a method based on local polynomial regression.

Another related line of research studies the construction of adaptive non-parametric confidence intervals. The work of Low (1997) showed that, in general, it is impossible to construct adaptive confidence bands simultaneously over different classes of Hölder functions; for recent results on the impossibility of adaptive confidence intervals, see Armstrong and Kolesár (2018) and references therein. Following that work, several studies have focused on identifying conditions under which adaptive confidence band construction is feasible. A well-studied condition is the one of self-similarity, which was first used in this context by Picard and Tribouley (2000) using wavelet methods for point-wise purposes. Later on, self-similarity was also used by Giné and Nickl (2010) to construct confidence bounds over finite intervals. Aside
from these two works, the self-similarity condition has also been used in other applications, including
high-dimensional sparse signal estimation (Nickl and van de Geer 2013), binary regression (Mukherjee
and Sen 2018), and $L_p$-confidence sets (Nickl and Szabó 2016), to mention but a few.

## 2 Problem formulation

**Notations and definitions.** For a positive integer $m$ we denote by $[m]$ the set \{1, \ldots, m\}. An open cube $U \subset \mathbb{R}^d$ with side-length $l > 0$ is the set $U = (x_1, x_1 + l) \times \ldots \times (x_d, x_d + l)$ for some $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ with center $c(U) = x + \frac{1}{2}(l, \ldots, l)$. For any $x \in \mathbb{R}^d$ denote by $\Xi(x, l)$ the open cube whose center is $x$ and has side-length $l$. For an open cube $U$ define $\Theta(U, l') := \{x \in U : \Xi(x, l') \subseteq U\}$ to be the set of points inside $U$ such that the open cube centered around them with side-length $l'$ falls completely inside $U$. For any function $f$ and set of points $U \subset \mathbb{R}^d$, we denote by $f|_U$ the restriction of $f$ to $U$. Furthermore, let $\| f \|_U = \sup_{x \in U} f(x)$.

For an integer $l \geq 0$, let $B_l := \{B_m, m = 1, \ldots, 2^d\}$ to be a re-indexed collection of the open cubes

$$B_m = B_m := \left\{ x \in [0, 1]^d : \frac{m_i - 1}{2^l} < x_i < \frac{m_i}{2^l}, i \in [d] \right\},$$

for $m = (m_1, \ldots, m_d)$ with $m_i \in [2^l]$.

For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we define $|s| = \sum_{i=1}^d s_i$, $s! = s_1! \ldots s_d!$, $x^s = x_1^{s_1} \ldots x_d^{s_d}$, and $||x|| = (x_1^2 + \ldots + x_d^2)^{1/2}$. Let $D^s$ denote the differential operator $D^s := \frac{\partial^{s_1 + \ldots + s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$.

Let $\beta > 0$. Denote by $\lfloor \beta \rfloor$ the maximal integer that is strictly less than $\beta$, e.g., $\lfloor 1 \rfloor = 0$. For any $x \in \mathbb{R}^d$ and any $\lfloor \beta \rfloor$ times continuously differentiable function $g$ on $\mathbb{R}^d$, we denote by $g_x$ its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point $x$:

$$g_x(x') := \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(x - x')^s}{s!} D^s g(x).$$

Finally, by log, we denote the natural logarithm. For an event $\mathcal{A}$ we denote by $\overline{\mathcal{A}}$ the complement of $\mathcal{A}$. For any set $S \subset \mathbb{R}^d$, let $\lambda[S]$ be the Lebesgue measure of $S$.

**Reward and feedback structure.** Let $\mathcal{K} = \{1, 2\}$ be a set of arms (actions) and let $\mathcal{T} = \{1, \ldots, T\}$ denote a sequence of decision epochs. At each time period $t \in \mathcal{T}$, a decision maker observes a context $X_t \in \mathcal{X} = [0, 1]^d$ that is realized according to an unknown distribution $P_X$, and then selects one of the 2
arms. When selecting an arm \( k \in \mathcal{K} \) at time \( t \in \mathcal{T} \), a reward

\[
Y_{k,t} \sim P_{Y|X}^{(k)}
\]
is realized and observed such that \( Y_{k,t} \in \{0, 1\} \), where \( P_{Y|X}^{(k)} \) is the distribution of the outcome conditional on context \( X_t \) and selecting arm \( k \). Equivalently, the outcomes \( Y_{k,t} \) may be expressed as follows:

\[
Y_{k,t} = f_k(X_t) + \epsilon_{k,t},
\]
where \( f_k(X_t) = \mathbb{E}[Y_{k,t} | X_t] \) and \( \epsilon_{k,t} \) is a random variable such that \( \mathbb{E} [\epsilon_{k,t} | X_t] = 0 \). The conditional distributions \( P_{Y|X}^{(k)} \) and the functions \( f_k \) are assumed to be unknown.

**Admissible policies and performance.** Let \( U \) be a random variable defined over a probability space \((\mathbb{U}, \mathcal{U}, \mathbb{P})\). Let \( \pi_t : \mathcal{X}^t \times [0, 1]^{t-1} \times \mathbb{U} \to \mathcal{K} \) for \( t = 1, 2, 3, \ldots \) be a sequence of measurable functions (with some abuse of notation we also denote the action at time \( t \) by \( \pi_t \in \mathcal{K} \)) given by

\[
\pi_t = \begin{cases} 
\pi_1(X_1, U) & t = 1, \\
\pi_t(X_t, \ldots, X_1, Y_{t-1}, \ldots, Y_1, U) & t = 2, 3, \ldots
\end{cases}
\]
The mappings \( \{\pi_t : t = 1, \ldots, T\} \), together with the distribution \( \mathbb{P} \) define the class of admissible policies, denoted by \( \Pi \). For any problem instance \( P = (P_X, P_{Y|X}^{(1)}, P_{Y|X}^{(2)}) \) let \( \pi^*(P) = (\pi^*_t(P), t = 1, 2, \ldots) \) denote the oracle rule, which at each period \( t \), under knowledge of the problem instance \( P \) (including the functions \( f_k \)), prescribes the best arm given the realized context \( X_t \):

\[
\pi^*_t(P) = \arg \max_{k \in \mathcal{K}} f_k(X_t), \quad t = 1, 2, \ldots
\]
Furthermore, the performance of a policy \( \pi \) is measured by the regret relative to the oracle performance:

\[
\mathcal{R}^\pi(P; T) := \mathbb{E}^\pi \left[ \sum_{t=1}^{T} f_{\pi^*_t(P)}(X_t) - f_{\pi_t}(X_t) \right].
\]

### 2.1 General model assumptions

We next detail the main model assumptions. Notably, all these assumptions are conventional in the literature of non-parametric contextual MAB (see, e.g., [Perchet and Rigollet 2013](#)).

**Covariate distribution.** The first assumption is regarding the distribution of the contexts. We will

require this distribution to be bounded from above and also away from zero. This way, we can make sure that in every given region over the covariate space, we can collect enough number of samples in order to be able to estimate the payoff functions \( f_k \).

**Assumption 1.** The distribution \( P_X \) is equivalent to the Lebesgue measure on \( \mathcal{X} \), that is, there exist constants \( 0 < \rho \leq \bar{\rho} \) such that for all \( x \in \mathcal{X} \),

\[
\rho \leq p_X(x) \leq \bar{\rho},
\]

where \( p_X \) is the density of the distribution \( P_X \).

**Smoothness.** We also need to make an assumption regarding how abrupt the payoff functions \( f_k \) can change in any given region. Intuitively, the smoother these functions are, the easier it is to estimate them. We formalize this notion using Hölder smoothness that is a classic definition of smoothness in the literature of non-parametric methods.

**Definition 2.1.** The Hölder class of functions \( \mathcal{H}_U(\beta, L) \) for the parameters \( \beta > 0 \) and \( L > 0 \) in an open cube \( U \subseteq (0, 1)^d \) is defined as the set of functions \( f : U \to \mathbb{R} \) that are \( \lceil \beta \rceil \) times continuously differentiable, and for any \( x, x' \in U \), satisfy the following inequality

\[
|f(x') - f_x(x')| \leq L\|x - x'\|_\infty.\]

Furthermore, let \( \mathcal{H}_U(\beta) := \bigcup_{0 \leq L < \infty} \mathcal{H}_U(\beta, L) \). We drop the indication \( U \) whenever \( U = (0, 1)^d \).

**Assumption 2.** The payoff functions \( f_k, \ k \in \mathcal{K} \), belong to the Hölder class of functions \( \mathcal{H}(\beta, L) \) for some \( L > 0 \) and \( \beta \in [\underline{\beta}, \bar{\beta}] \) with \( 0 < \underline{\beta} \leq 1 \).

**Margin condition.** The next assumption aims to capture the interplay between the payoff functions and the covariate distribution. A key factor in determining the complexity of our problem is the mass of covariates near the decision boundary. Intuitively, the less concentrated the contexts are near the boundary, the easier the problem is since distinguishing the optimal decision away from the boundary is relatively a simpler task and it is near the boundary that an agent will potentially make more mistakes.

**Assumption 3.** There exist some \( \alpha > 0 \) and \( C_0 > 0 \) such that

\[
P_X \{0 < |f_1(X) - f_2(X)| \leq \delta \} \leq C_0 \delta^\alpha \quad \forall \delta > 0.
\]
This assumption is known as the margin condition in the literature. One can see that the larger the parameter $\alpha$, the faster the covariate mass shrinks near the boundary (where it is harder to distinguish the optimal arm), which means as $\alpha$ grows the hardness of the problem reduces. The above three assumptions characterize the general class of problems that we are interested in.

**Definition 2.2.** For any $\beta \geq 0$ and $\alpha \geq 0$, we denote by $\mathcal{P}(\beta, \alpha, d) = \mathcal{P}(\beta, L, \alpha, C_0, \bar{\rho}, \rho, d)$ the class of problems that satisfy Assumption [1] for some $\bar{\rho} \geq \rho > 0$, Assumption [2] for $\beta$ and some $L > 0$, and Assumption [3] for $\alpha$ and some $C_0 > 0$.

It is worth noting that the smoothness condition and the margin condition are related to each other. Essentially, the smoothness of payoff functions determines how abrupt they can change near the decision boundary, which in turn, affects the mass of contexts in that region. That is, when payoff functions are smooth (the parameter $\beta$ is large), they cannot move away from each other too fast, hence, resulting in more mass of contexts near the decision boundary (the parameter $\alpha$ is small). This relationship is stated concretely in the following proposition, which is a simple extension of Proposition 3.1 in Perchet and Rigollet (2013).

**Proposition 2.3.** Assume that Assumption [2] holds with parameters $(\beta, L)$, and that Assumption [3] holds with parameter $\alpha$. Then, the following statements hold:

1. If $\alpha(1 \wedge \beta) > 1$ then, a given arm is always or never optimal; the oracle policy $\pi^*$ dictates pulling only one of the arms all the time;

2. If $\alpha(1 \wedge \beta) \leq 1$ then, there exits problem instances in $\mathcal{P}(\beta, \alpha, d)$ with non-trivial oracle policies.

Based on this proposition, the interesting case is when $0 < \alpha \leq \frac{1}{1 \wedge \beta}$, which is the range that we will consider in the rest of the paper.

## 3 Impossibility of adapting to smoothness

In this section, we discuss the possibility of adapting to the smoothness of payoff functions under the general model assumptions stated in the previous section. As background, suppose that it is a priori known that the problem instance belongs to $\mathcal{P}(\beta, \alpha, d)$. Then, the minimax regret achievable over $T$ time periods scales as

$$\inf_{\pi \in \Pi} \sup \{ R^\pi(\mathcal{P}; T) : \mathcal{P} \in \mathcal{P}(\beta, \alpha, d) \} = \Theta \left( T^{\zeta(\beta, \alpha, d)} \right), \quad \zeta(\beta, \alpha, d) = 1 - \frac{\beta(1 + \alpha)}{2\beta + d}. \quad (3.1)$$
This result was first established by Rigollet and Zeevi (2010) and Perchet and Rigollet (2013) in the case \( \beta \leq 1 \), and by Hu et al. (2019) in the case with \( \beta > 1 \) under further assumptions on the regularity of the decision regions.

The goal of an adaptive bandit would be to achieve the rate of convergence \((3.1)\) without prior knowledge of the best smoothness parameter \( \beta \) characterizing the payoff functions \( f_k \). Our first result, however, shows that this is impossible in the following sense. Suppose that, given some smoothness parameters \( 0 < \beta < \gamma \leq \frac{1}{2} \), it is known that \( P \in \mathcal{P}(\beta, \alpha, 1) \), but unknown whether or not \( P \in \mathcal{P}(\gamma, \alpha, 1) \). Then, for any policy \( \pi \) that achieves the optimal rate \((3.1)\) over the smoother class \( \mathcal{P}(\gamma, \alpha, 1) \), there exists some problem instance \( P \in \mathcal{P}(\beta, \alpha, 1) \) for which \( R_\pi(P; T) = \omega(T^{\zeta(\beta, \alpha, d)}) \). In other words any policy \( \pi \) that achieves optimal rates of regret uniformly over the smoother class of functions \( \mathcal{P}(\gamma, \alpha, 1) \) cannot also guarantee optimal regret rates uniformly over the more general class \( \mathcal{P}(\beta, \alpha, 1) \).

**Theorem 3.1.** Fix two H"older exponents \( 0 < \gamma \leq \frac{1}{2} \), \( 0 < \beta < \frac{\gamma}{2-\alpha\gamma} \wedge \frac{\alpha\gamma^2}{1-4\gamma} \), some parameter \( 0 < \alpha \leq \frac{1}{7} \) and let \( d = 1 \). For any admissible policy \( \pi \in \Pi \) that achieves rate-optimal regret over \( \mathcal{P}(\gamma, \alpha, 1) \), we have

\[
\sup \{ R_\pi(P; T) : P \in \mathcal{P}(\beta, \alpha, 1) \} \geq CT^{1-\frac{\alpha\gamma^2+\gamma-\beta}{\alpha(2\beta\gamma+\gamma-\beta)}} \left[ \sup \{ R_\pi(P; T) : P \in \mathcal{P}(\gamma, \alpha, 1) \} \right]^{-\frac{\alpha\gamma^2+\gamma-\beta}{2\beta\gamma+\gamma-\beta}}
\]

for some constant \( C > 0 \).

Theorem 3.1 establishes a lower bound on the achievable performance over a class of problems as a function of the performance over another class of problems with smoother payoff functions. This lower bound depends on the smoothness parameters of the two considered classes of payoff functions and also the margin condition parameter.

**Key ideas in the proof.** The proof of Theorem 3.1 adapts to our framework ideas of identifying a worst-case nature “strategy.” While the full proof is deferred to the appendix, we next describe its key ideas using the special case of \( \alpha = \frac{1}{7} \); the construction of the worst-case instance in this special case is depicted in Figure 2.

Fix a parameter \( \Delta \leq \frac{1}{4} \). First, consider a problem instance in \( \mathcal{P}(\gamma, \alpha, 1) \) such that the first arm’s payoff function is \( \frac{1}{2} \) everywhere except for the interval \([0, 2\Delta^\frac{3}{2}]\), where it has a “downward bump” and reaches its minimum, \( \frac{1}{2} - \Delta \) as depicted in Figure 2, and the second arm’s payoff function is \( \frac{1}{2} \) everywhere. Furthermore, for each \( 1 \leq m \leq M := \left| \frac{1}{2} - \frac{1}{2} \right| \), consider a problem instance in \( \mathcal{P}(\beta, \alpha, 1) \) such that the payoff functions are equal to the aforementioned payoff functions everywhere except for the interval \( I_m := [4(m-1)\Delta^\frac{3}{2}, 4m\Delta^\frac{3}{2}] \), where the first arm’s payoff function has an “upward bump” and reaches its maximum, \( \frac{1}{2} + \Delta \) as depicted in Figure 2, that is, for the problem \( m \), the first arm is optimal over some
section of \( I_m \). Due to the assumption that the regret of the policy \( \pi \) over \( \mathcal{P}(\gamma, \alpha, 1) \) is upper bounded by \( \sup \{ \mathcal{R}^\pi(\mathcal{P}; T) : \mathcal{P} \in \mathcal{P}(\gamma, \alpha, 1) \} \), the number of times the policy pulls arm 1 inside at least one of the intervals \( I_m \) has to be “small.” Let \( m^* \) be the index of one such interval. Using this observation and by choosing a specific value for \( \Delta \), we show that with a strictly positive probability, the policy will not be able to “distinguish” between the problem instance in \( \mathcal{P}(\gamma, \alpha, 1) \) and the problem \( m^* \) in \( \mathcal{P}(\beta, \alpha, 1) \). This causes the policy not to pull arm 1 inside the corresponding interval \( I_m^* \), that contains a section over which, the first arm is optimal for the problem \( m^* \); hence, guaranteeing the regret lower bound stated in the statement of the theorem.

3.1 Discussion

Through Theorem 3.1, we can, in fact, show that in general, there exists no single policy that can simultaneously achieve rate optimal performance over different classes of problems characterized by different smoothness. Suppose policy \( \pi \) achieves rate-optimal performance over \( \mathcal{P}(\gamma, \alpha, 1) \), that is, \( \sup \{ \mathcal{R}^\pi(\mathcal{P}; T) : \mathcal{P} \in \mathcal{P}(\gamma, \alpha, 1) \} = C T^{\zeta(\gamma, \alpha, 1)} \) for some constant \( C > 0 \). Consider some \( 0 < \beta < \frac{\gamma}{2-\alpha \gamma} \land \frac{\alpha \gamma^2}{1-\frac{\gamma}{2}} \), and assume \( \gamma \leq \frac{1}{2} \). Then, Theorem 3.1 implies that there exists a problem instance in \( \mathcal{P}(\beta, \alpha, 1) \) for which

\[
\mathcal{R}^\pi(\mathcal{P}; T) \geq C T^1 \frac{\alpha^2 - 2\beta^2 + 2\beta^2 + 2\beta + \gamma - \gamma}{(\gamma + 2\beta + \gamma - \beta)(2\beta + 1)},
\]

for some constant \( C > 0 \). Next, we provide an example, where this lower bound is greater than the optimal regret rate \( T^{\zeta(\beta, \alpha, 1)} \) over \( \mathcal{P}(\beta, \alpha, 1) \).

Example 1. Let \( \gamma = 0.15 \), \( \beta = \frac{\gamma}{2} \), and \( \alpha = \frac{99}{100 \gamma} \). Then, the optimal regret rate with the knowledge of smoothness over \( \mathcal{P}(\beta, \alpha, 1) \) is \( \mathcal{O}(T^{0.504348}) \), while Theorem 3.1 establishes a lower bound of order \( \Omega(T^{0.57966}) \) if the policy \( \pi \) achieves rate-optimal performance over \( \mathcal{P}(\gamma, \alpha, 1) \).
Despite the impossibility of adapting to unknown smoothness that is implied by Theorem 3.1 in the next section we will leverage a condition that allows adaptivity to the smoothness of payoff functions.

4 Adaptivity to global smoothness

In this section, we first introduce in §4.1 the global self-similarity condition. In §4.2 we briefly review local polynomial regression, which is the key estimation tool in our proposed policy. Then, in §4.3 we describe our policy and establish that it guarantees near-optimal performance without prior knowledge of the best smoothness that characterizes payoff functions by leveraging the self-similarity notion.

4.1 A sufficient condition for adapting to global smoothness

We leverage the global self-similarity condition that appears in the literature on non-parametric confidence bands (see, e.g., Giné and Nickl 2010 and Picard and Tribouley 2000) and tailor it to our setting. For any function \( f \), and non-negative integers \( l \) and \( p \), define \( \Gamma_{p}^{l} f(\cdot; U) \) to be the \( L_2(\mathbb{P}_X) \)-projection of the function \( f \) to the class of polynomial functions of degree at most \( p \) over the open cube \( U \). More precisely, for any \( x \in U \):

\[
\Gamma_{p}^{l} f(x; U) := g(x), \quad \text{s.t. } g = \arg \min_{q \in \text{Poly}(p)} \int_{U} |f(u) - q(u)|^2 K \left( \frac{x - u}{h} \right) p_X(u \mid U) du,
\]

where \( K(\cdot) = 1 \{ ||\cdot||_{\infty} \leq 1 \} \), \( h = 2^{-l} \), and \( \text{Poly}(p) \) is the class of polynomials of degree at most \( p \). The next lemma provides a closed form expression for \( \Gamma_{l}^{p} f(\cdot; U) \) along with some properties of this projection.

Lemma 4.1. Fix a positive integer \( l \), an open cube \( U \) of side-length \( 2^{-l'} \), \( l' \in \mathbb{R}_+ \), and a point \( x \in U \) and let \( K(\cdot) = 1 \{ ||\cdot||_{\infty} \leq 1 \} \) and \( h = 2^{-l} \). Let \( \mu_0, \kappa_0, \) and \( L_0 \) be some constants that only depend on \( p, \rho, \bar{\rho} \) (introduced in Assumption 1), and \( d \). The following statements hold:

a. \( \Gamma_{l}^{p} f(x; U) = R^\top (0) B^{-1} W \) where we define the vector \( R(u) := (u^s)_{|s| \leq p} \), the matrix \( B := (B_{s_1, s_2})_{|s_1|, |s_2| \leq p} \) and the vector \( W := (W_s)_{|s| \leq p} \) with elements

\[
B_{s_1, s_2} := \int_{\mathbb{R}^d} u^{s_1 + s_2} K(u) p_X(x + hu \mid U) du, \quad W_s := \int_{\mathbb{R}^d} u^s f(x + hu) K(u) p_X(x + hu \mid U) du;
\]

b. \( \lambda_{\text{min}}(B) \geq \mu_0 2^{d l'} \);

c. \( |\Gamma_{l}^{p} f(x; U) - \Gamma_{l}^{p} f(\hat{x}; U)| \leq \kappa_0 h^{-1} \| \hat{x} - x \|_{\infty} \) for all \( x, \hat{x} \in B \);
d. If \( f \in \mathcal{H}(\beta, L) \) for \( 0 < \beta \leq p + 1 \) then, \(|\Gamma^p f(x; U) - f(x)| \leq L_0 h^\beta\) for all \( x \in B \).

Next, we formalize the notion of global self-similarity using the projection \( \Gamma^p f \).

**Definition 4.2 (Global self-similarity).** A function is \( f \) is said to be globally self-similar if \( f \in \mathcal{H}_U(\beta) \) for some \( \beta > 0 \), and there exist some positive integer \( l_0 \) and some constants constant \( b > 0 \) such that for any \( l \geq l_0 \) and integer \( \lfloor \beta \rfloor \leq p \leq \lfloor \bar{\beta} \rfloor \),

\[
\max_{B \in B_i} \| \Gamma^p f(x; B) - f(x) \|_B \geq b 2^{-l}. 
\]

**Assumption 4.** All the payoff functions \( f_k, k \in K \), are globally self-similar.

As we will see in the next subsection, local polynomial regression estimation of any function \( f \) cannot deviate from the projection \( \Gamma^p f \). That is, the above assumption together with Part d of Lemma 4.4 states that for a Hölder smooth function, not only is the bias of estimation bounded from above, but also it cannot shrink fast globally. As we will see, when this condition holds, one may be able to estimate the smoothness of the function by examining the estimation bias against its variance over the unit cube. Consequently, we will utilize this estimated smoothness in order to achieve optimal regret rate up to poly-logarithmic terms.

### 4.2 A review of local polynomial regression

Since our proposed policy is based on local polynomial regression, in this subsection, we shortly review the local polynomial regression method based on the analysis in Audibert and Tsybakov (2007). Let \( D = \{(X_i, Y_i)\}_{i=1}^n \) be a set of \( n \) i.i.d. data pairs \((X_i, Y_i) \in \mathcal{X} \times \mathbb{R}\), distributed according to a joint distribution \( P \). Denote by \( \mu \) the marginal density of \( X_i \)'s and define the regression function \( \eta(x) := \mathbb{E}[Y|X = x] \). In order to estimate the value of the function \( \eta \) at any point \( x \in \mathcal{X} \) we deploy local polynomial regression which is defined as follows:

**Definition 4.3.** Fix a set of pairs \( D = \{(X_i, Y_i)\}_{i=1}^n \), a point \( x \in \mathbb{R}^d \), a bandwidth \( h > 0 \), an integer \( p > 0 \) and a kernel function \( K : \mathbb{R}^d \rightarrow \mathbb{R}_+ \). Define by \( \hat{\theta}_x(u; D, h, p) = \sum_{|s| \leq p} \xi_s u^s \) a polynomial of degree \( p \) on \( \mathbb{R}^d \) that minimizes

\[
\sum_{i=1}^n \left( Y_i - \hat{\theta}_x(X_i - x; D, h, p) \right)^2 K \left( \frac{X_i - x}{h} \right). \tag{4.2}
\]

The local polynomial estimator \( \hat{\eta}^{LP}(x; D, h, p) \) of the value \( \eta(x) \) of the regression function \( f \) at point \( x \) is defined by: \( \hat{\eta}^{LP}(x; D, h, p) := \hat{\theta}_x(0; D, h, p) \) if the above expression has a unique minimizer, and \( \hat{\eta}^{LP}(x; D, h, p) := 0 \) otherwise.
Define the matrix \( Q := (Q_{s_1,s_2})_{|s_1|,|s_2| \leq p} \) and the vector \( V := (V_s)_{s \leq p} \) with the elements

\[
Q_{s_1,s_2} := \sum_{i=1}^{n} (X_i - x)^{s_1+s_2} K \left( \frac{X_i - x}{h} \right), \quad V_s := \sum_{i=1}^{n} Y_i (X_i - x)^s K \left( \frac{X_i - x}{h} \right).
\]

Also, define the matrix \( U := (u^s)_{|s| \leq p} \). The next result from Audibert and Tsybakov (2007) provides a closed-form expression for local polynomial regression at any arbitrary point.

**Lemma 4.4** (Audibert and Tsybakov 2007, Proposition 2.1). If the matrix \( Q \) is positive definite, there exists a polynomial on \( \mathbb{R}^d \) of degree \( p \) minimizing (4.2). Its vector of coefficients is given by \( \xi = Q^{-1}V \) and the corresponding local polynomial regression function at point \( x \) is given by

\[
\hat{\eta}^{LP}(x; \mathcal{D}, h, p) = U(0)^\top Q^{-1}V = \sum_{i=1}^{n} Y_i K \left( \frac{X_i - x}{h} \right) U(0)^\top Q^{-1}U(X_i - x).
\]

In what follows, the kernel function will be the unit rectangle kernel \( K(\cdot) = 1 \{ \|\cdot\|_\infty \leq 1 \} \). The following simple extension of Theorem 3.2 in Audibert and Tsybakov (2007) will be one of the main tools to bound our estimation error in our proposed policy.

**Proposition 4.5.** Let \( \mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n} \) be a set of \( n \) i.i.d pairs \( (X_i, Y_i) \in \mathcal{X} \times \mathbb{R} \). If the marginal density \( \mu \) of \( X_i \)'s satisfies \( \underline{\mu} \leq \mu(x) \leq \bar{\mu} \) for some \( 0 < \underline{\mu} \leq \bar{\mu} \) with a support \( \mathcal{X} \) that is a closed cube in \( \mathbb{R}^d \), and the function \( \eta \) belongs to the Hölder class of functions \( \mathcal{H}_X(\beta, L) \) for some \( \beta, L > 0 \) then, there exist constants \( C_1, C_2, C_3 > 0 \) such that for any \( 0 < h < r_0 \), any \( C_3 h^\beta < \delta \), any \( n \geq 1 \) and the kernel function \( K(\cdot) = 1 \{ \|\cdot\|_\infty \leq 1 \} \), the local polynomial estimator \( \hat{\eta}^{LP}(x; \mathcal{D}, h, p) \) satisfies

\[
|\hat{\eta}^{LP}(x; \mathcal{D}, h, p) - \eta(x)| \leq \delta
\]

with probability at least \( 1 - C_1 \exp \left(- C_2 nh^d \bar{\mu}^2 \delta^{-2}\right) \) for all \( x \in \mathcal{X} \). The constants \( C_1, C_2, C_3 \) depend only on \( p, d, L, c_0, \) and \( r_0 \).

The next proposition states that local polynomial regression estimation of a function inside an open cube cannot deviate from the \( L_2(P) \)-projection of that function inside the cube.

**Proposition 4.6.** Fix an open cube \( U \subseteq (0,1)^d \) with side-length \( 2^{-l'} \), \( l' \in \mathbb{R}_+ \). Let \( \mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n} \) be a set of \( n \) i.i.d pairs \( (X_i, Y_i) \in U \times \mathbb{R} \). If the marginal density \( \mu \) of \( X_i \)'s satisfies \( \mu(\cdot) = p_X(\cdot|U) \), where \( p_X \) is the density of a distribution \( P_X \) that satisfies Assumption \( \mathbb{A} \) then, there exist constants \( C_4, C_5, C_6 > 0 \) such that for any \( \delta < C_6 \), any \( n \geq 1 \), \( h = 2^{-l}, l \geq l' \), and the kernel function \( K(\cdot) = 1 \{ \|\cdot\|_\infty \leq 1 \} \), the
local polynomial estimator $\hat{\eta}^{\text{LP}}(x; D, h, p)$ satisfies

$$|\hat{\eta}^{\text{LP}}(x; D, h, p) - \Gamma_p \eta(x; U)| \leq \delta$$

with probability at least $1 - C_4 \exp\left(-C_5 n 2^{d(l'-l)} \delta^2\right)$ for all $x \in U$. The constants $C_4, C_5, C_6$ depend only on $p, \bar{\rho}, \rho, d$.

4.3 A proposed policy

Our proposed policy integrates a smoothness estimation sub-routine with a rate-optimal policy to build a rate-optimal adaptive policy. The smoothness estimation sub-routine consists of three consecutive steps: collecting samples at different regions of the context space, estimating the payoff functions based on the collected samples, and running a statistical test over the estimated functions. The policy will repeat these steps until it is terminated. At the end, it will estimate the global smoothness of the payoff functions based on the results of the statistical tests. Next, we will explain each of these steps separately.

4.3.1 Estimating smoothness

**Sampling.** In the SACB policy detailed in Algorithm 1, we will consider the partition of the unit cube corresponding to $B_l$ with $l = \left\lceil \frac{(\beta + d - 1) \log_2 T}{2(\beta + d)^2} \right\rceil$. For each open cube $B \in B_l$, we collect samples for both arms in multiple rounds. Define the maximum round index as follows:

$$\bar{r} := \bar{r}(T, d, \bar{\beta}, \beta) := \left\lceil 2l \beta + \frac{2d}{\beta} + 4 \log_2 \log T \right\rceil.$$

At every round $r \in [\bar{r}]$, $2^r$ samples are collected for each arm by alternating between them every time a context falls into $B$. If for some $B \in B_l$, we reach $\bar{r}$ before the SACB policy is terminated we simply pull arm 1 every time a context falls into $B$.

**Estimation.** Denote by $X_{k,1}(B), X_{k,2}(B), \ldots$ and $Y_{k,1}(B), Y_{k,2}(B), \ldots$ the successive covariates and outcomes when arm $k$ is pulled in $B$ at round $r$, respectively. Denote by $D_{k}^{(B,r)} := \left\{ (X_{k,\tau}(B), Y_{k,\tau}(B)) \right\}_{\tau=1}^{2^r}$ the corresponding set of pairs. Define the two bandwidth exponents:

$$j_1^{(B)} := l, \quad j_2^{(B)} := l + \left\lceil \frac{1}{\beta} \log_2 \log T \right\rceil.$$
Let \( \hat{l} := \lceil \frac{\beta l}{2} + \frac{\log_2 \log T}{\beta} \rceil \lor \lceil (1 + \tilde{\beta})l + \log_2 \log T \rceil \). For every bin \( B \) define the mesh points:

\[
\mathcal{M}^{(B)} := \left\{ x = \left( \frac{m_1}{2^l}, \ldots, \frac{m_d}{2^l} \right) : x \in B, m_i \in \left[ 2^{\hat{l}} \right] \text{ for } i \in [d] \right\}.
\]

For every mesh point \( x \) in \( \mathcal{M}^{(B)} \), we form two separate estimates of the payoff functions using local polynomial regression of degree \( \lceil \tilde{\beta} \rceil \):

\[
\hat{f}_k^{(B,r)}(x; j) := \hat{f}_k(x; D_k^{(B,r)}, 2^{-j}, \lceil \tilde{\beta} \rceil), \quad j \in \{ j_1^{(B)} \cup j_2^{(B)} \}.
\]

**Statistical test.** At the end of each sampling round \( r \), we check if for any of the open cubes \( B \in \mathcal{B}_l \), the difference between the estimation using the two bandwidths exponents \( j_1^{(B)} \) and \( j_2^{(B)} \) exceeds a pre-determined threshold. Namely, we check if the following holds true:

\[
\sup_{k \in \mathcal{K}, x \in \mathcal{M}^{(B)}} \left| \hat{f}_k^{(B,r)}(x; j_1^{(B)}) - \hat{f}_k^{(B,r)}(x; j_2^{(B)}) \right| \geq \gamma \frac{(\log T)^{\frac{d}{2} + \frac{1}{2}}}{2r/2}, \tag{4.3}
\]

where \( \gamma \) is a tuning parameter. Broadly speaking, the left hand side of \( (4.3) \) is dominated by two terms: the estimation bias of \( \hat{f}_k^{(B,r)}(x; j_1^{(B)}) \) and the standard deviation of \( \hat{f}_k^{(B,r)}(x; j_2^{(B)}) \). The reason is that the former has a larger bandwidth, which results in a potentially larger bias, and the latter is an estimation based on less samples on average, which results in a larger standard deviation. Furthermore, the right hand side of \( (4.3) \) is proportional to standard deviation of the estimate \( \hat{f}_k^{(B,r)}(x; j_2^{(B)}) \). That is, by examining \( (4.3) \), we are detecting the number of samples that is required for the estimation bias of \( \hat{f}_k^{(B,r)}(x; j_1^{(B)}) \) to dominate the standard deviation of \( \hat{f}_k^{(B,r)}(x; j_2^{(B)}) \), which, as we will see, is dependent on how smooth the payoff functions behave. This dependence will allow us to infer the smoothness of payoff functions with a good precision with high probability. Denote by \( r_{\text{last}}^{(B)} \) the smallest round for which \( (4.3) \) holds in the open cube \( B \). If this inequality never happens in \( B \), we simply set \( r_{\text{last}}^{(B)} = \bar{r} \). The next proposition provides a high-probability lower bound for \( r_{\text{last}}^{(B)} \) for all the open cubes \( B \in \mathcal{B}_l \).

**Proposition 4.7.** Under Assumption \( \mathbb{A} \), there exist constants \( C_r, C_\gamma, C_8, C_9 \) such that the probability that

\[
r_{\text{last}}^{(B)} < C_r + 2l\beta + \left( \frac{d}{2} + 1 \right) \log_2 \log T
\]

for some \( B \in \mathcal{B}_l \), is less than \( C_\gamma 2^{dL} (\log T)^{\frac{d}{2}} T^{-\gamma^2 C_8 + C_9} \), where the constants \( C_\gamma, C_8, \) and \( C_9 \) depend on only \( \beta, \tilde{\beta}, L, \rho, \tilde{\rho}, \) and \( d \), and \( C_r \) depends on only \( \beta, \tilde{\beta}, L, \rho, \) and \( \tilde{\rho} \).

**Key ideas in the proof.** The proof is based on the interpretation that we have provided after \( (4.3) \).
That is, since the payoff functions are in $\mathcal{H}(\beta, L)$, their estimation bias is bounded inside each open cube $B \in \mathcal{B}_l$, which implies that the left hand side of (4.3) is dominated by the standard deviation of the estimate $\hat{f}_k^{(B,r)}(x;j_2^{(B)})$, which is proportional to the right hand side of (4.3). Hence, when the number of samples is “small” in the sense stated in the proposition, (4.3) fails to hold with high probability.

The next proposition is a partial converse of the previous proposition and provides a high-probability upper bound for $\min_{B \in \mathcal{B}_l} r_{\text{last}}^{(B)}$.

**Proposition 4.8.** *Under Assumption 4, there exist some $B \in \mathcal{B}_l$ as well as some constants $C_r$, $C_{10}$, and $C_{11}$ such that the probability that

$$r_{\text{last}}^{(B)} > C_r + 2l\beta + \left(\frac{d}{\beta} + 3\right) \log_2 \log T$$

is less than $C_{10}T^{-\gamma^2C_{11}}$, where the constants $C_{10}$ and $C_{11}$ depend on only $\beta, \tilde{\beta}, L, \rho, \bar{\rho}$, and $d$, and $C_r$ depends on only $\beta, \tilde{\beta}, L, \rho, \bar{\rho}$.

**Key ideas in the proof.** The proof idea of this proposition is also based on the interpretation that we have provided after (4.3). That is, since the payoff functions are globally self-similar, the estimation bias of the estimate $\hat{f}_k^{(B,r)}(x;j_1^{(B)})$ remains “large” in at least one of the open cubes $B \in \mathcal{B}_l$ (even larger than that of the estimate $\hat{f}_k^{(B,r)}(x;j_2^{(B)})$, which implies that for that specific open cube, if the number of samples is “large” enough in the sense stated in the proposition then, the left hand side of (4.3) will be dominated by the aforementioned bias and will eventually exceed the right hand side of (4.3) with high probability.

Based on propositions 4.7 and 4.8 we estimate the global smoothness of the problem according to the following expression:

$$\hat{\beta}_{\text{SACB}} = \left[ \min_{B \in \mathcal{B}_l} r_{\text{last}}^{(B)} - \left(\frac{2d}{\beta} + 4\right) \log_2 \log T \right] / 2l.$$ 

Next, we characterize the accuracy by which the SACB policy estimates the global smoothness $\beta$.

**Corollary 4.9.** *Under Assumptions 3 and 4 and for large enough $T \geq T_0(\beta, \tilde{\beta}, L, \rho, \bar{\rho})$

$$P\left\{ \hat{\beta}_{\text{SACB}} \in \left[ \beta - \frac{3(2\beta + d)^2 \log_2 \log T}{(\beta + d - 1) \log_2 T}, \beta \right] \right\} \geq 1 - C_{12} (\log T)^\frac{d}{2} T^{-\gamma^2C_{13} + C_{14}},$$

where the constants $C_{12}, C_{13}$, and $C_{14}$ depend on only $\beta, \tilde{\beta}, L, \rho, \bar{\rho}$, and $d$.

Note that we choose the final estimate of the global smoothness such that it is less than $\beta$ with high probability. This is known as "undersmoothing" in the construction of confidence intervals; see, e.g.,
4.3.2 A rate-optimal policy

Let $\pi^{opt}$ be a contextual MAB policy that achieves rate-optimal regret when initialized with the right parameter $\beta$. In this step, we initialize $\pi^{opt}$ with $(T - T_{SACB}, \hat{\beta}_{SACB})$, where $T_{SACB}$ is the time step at which the smoothness estimation sub-routine is terminated, and run this policy for the rest of the time horizon. More precisely, if $\hat{\beta}_{SACB} \leq 1$ we deploy the Adaptively Binned Successive Elimination (ABSE) policy in Perchet and Rigollet (2013) as $\pi^{opt}$. The ABSE policy relies on the knowledge of $\beta$, and achieves the rate-optimal regret $O(T^{\zeta(\beta, \alpha, d)})$ for any problem instance with $0 < \beta \leq 1$. On the other hand, if $\hat{\beta}_{SACB} > 1$ we deploy the SmoothBandit policy in Hu et al. (2019) as $\pi^{opt}$. The SmoothBandit policy also relies on the knowledge of $\beta$ and achieves the near-optimal regret rate $O\left((\log T)^{\frac{2d + d}{2d}} T^{\zeta(\beta, \alpha, d)}\right)$ for any problem instance with $\beta \geq 1$ provided the following additional assumption regarding the regularity of decision regions holds:

**Assumption 5.** Let $Q_k := \{ x \in [0,1]^d : (-1)^{k-1}(f_1(x) - f_2(x)) \geq 0 \}, k \in K$, be the optimal decision regions. Then, each $Q_k$, is a non-empty $(c_0, r_0)$-regular set, where a Lebesgue measurable set $S$ is said to be $(c_0, r_0)$-regular if for all $x \in S$,

$$\lambda(S \cap \text{Ball}_2(x,r)) \geq c_0 \lambda(\text{Ball}_2(x,r)),$$

where $\text{Ball}_2(x,r)$ is the Euclidean ball of radius $r$ around $x$.

4.3.3 Pseudo-code

The pseudo-code that describes the SACB policy is presented below.
Algorithm 1: Smoothness-Adaptive Contextual Bandits (SACB)

1. **Input:** Time horizon $T$, minimum and maximum smoothness exponents $\beta$ and $\bar{\beta}$, and a tuning parameter $\gamma$

2. **Output:** The termination time $T_{\text{SACB}}$ and the estimated global smoothness $\hat{\beta}_{\text{SACB}}$

3. **Initialize:**
   - $l \leftarrow \lceil \frac{(\beta + d - 1) \log_2 T}{2(\beta + d)} \rceil$ and $N_k^{(B)}(t) \leftarrow 0$ for all $k \in \mathcal{K}$ and $B \in \mathcal{B}_l$

   for $t = 1, \ldots$ do

   5. Determine the bin in which the current covariate is located: $B \in \mathcal{B}_l$ s.t. $X_t \in B$

   6. Alternate between the arms: $\pi_t \leftarrow 1 + \mathbb{1} \left\{ N_1^{(B)} > N_2^{(B)} \right\}$

   7. Update the counters: $N_k^{(B)} \leftarrow N_k^{(B)} + \mathbb{1} \left\{ \pi_t = k \right\} \forall k \in \mathcal{K}$

   8. Receive and observe reward $Y_{\pi_t, t}$

   if $N_1^{(B)} + N_2^{(B)} \geq 2 \times 2^r(B)$ and $r(B) \leq \bar{r}$ then

   | if $\sup_{k \in \mathcal{K}, x \in \mathcal{M}(B)} \left| \hat{f}_k^{(B, r(B))}(x; j_1^{(B)}) - \hat{f}_k^{(B, r(B))}(x; j_2^{(B)}) \right| > \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{\bar{r}T} / 2}$ then

   | Record $r_{\text{last}}^{(B)}$: $r_{\text{last}}^{(B)} \leftarrow r(B)$

   | Collect double number of samples in the next round: $r(B) \leftarrow r(B) + 1$; Reset the counters: $N_k^{(B)} \leftarrow 0 \forall k \in \mathcal{K}$

   else if $r(B) > \bar{r}$ then

   | Pull arm 1: $\pi_t = 1$

   if $r(B') > \bar{r} \forall B' \in \mathcal{B}_l$ then

   | Record the termination time: $T_{\text{SACB}} \leftarrow t$

   break

18. Estimate the global smoothness: $\hat{\beta}_{\text{SACB}} \leftarrow \left\lfloor \min_{B \in \mathcal{B}_l} \left( r_{\text{last}}^{(B)} - \frac{(2d)}{2} + 4 \log_2 \log T \right) / 2l \right\rfloor$

19. if $\hat{\beta}_{\text{SACB}} \leq 1$ then

   $\pi_{\text{opt}} \leftarrow \text{ABSE}$

21. else

   Initialize $\pi_{\text{opt}}$ with $\left( T - T_{\text{SACB}}, \hat{\beta}_{\text{SACB}} \right)$

24. for $t = T_{\text{SACB}} + 1, \ldots, T$ do

   $\pi_t \leftarrow \pi_{\text{opt}}(X_t)$

The next theorem characterizes the regret rate of the SACB policy for the cases when $\beta \leq 1$ and $\beta \geq 1$.

**Theorem 4.10.** Suppose policy $\pi$ detailed in Algorithm 1 is run, and consider a problem instance $P \in \mathcal{P}(\beta, \alpha, d)$ with $\underline{\beta} \leq \beta \leq \overline{\beta}$ such that the payoff functions associated with $P$ satisfy the global self-similarity condition in Assumption 4. Then, the followings hold for large enough tuning parameter $\gamma \geq \gamma_0(\underline{\beta}, \overline{\beta}, L, \rho, \bar{\rho}, d)$ and time horizon $T \geq T_0(\underline{\beta}, \overline{\beta}, L, \rho, \bar{\rho}, d)$:
a. If $\beta \leq 1$ we have
\[
R^\pi(P; T) = O \left( T^{\zeta(\beta, \alpha, d)} \left( \log T \cdot \frac{3d(\alpha + 1)(2\beta + d)^2}{(2\beta + d)(2\beta + d - 1)} \right) \right);
\]

b. If $\beta > 1$ and the decision regions associated with $P$ satisfy the regularity condition in Assumption 5 we have
\[
R^\pi(P; T) = O \left( T^{\zeta(\beta, \alpha, d)} \left( \log T \cdot \frac{3d(\alpha + 1)(2\beta + d)^2}{(2\beta + d)(2\beta + d - 1)^2} + \frac{\beta + d}{\beta} \right) \right).
\]

**Key ideas of the proof.** In the proof, we show that the regret incurred due to running the sub-routine that estimates the smoothness is negligible compared to the optimal regret rate. Furthermore, we show that the regret incurred due to running $\pi^{\text{opt}} = \pi^{\text{opt}}(T_{S\text{SACB}}, \hat{\beta}_{S\text{SACB}})$ is of near-optimal rate times some poly-logarithmic term that stems from the error in the estimate $\hat{\beta}_{S\text{SACB}}$.

## 5 Local smoothness

In the previous section we introduced a method to adapt to the global smoothness of a contextual MAB problem under the assumption of global self-similarity. However, classifying the hardness of the problem, and predating policy design on a single global smoothness parameter might be associated with fundamental inefficiency. For example, consider the one-dimensional problem instance shown in Figure 3 in which there are regions over the covariate space where payoff functions behave differently in terms of smoothness. The global smoothness parameter of this problem is dictated by how smooth the payoff functions behave in the roughest region. That is, in trying to adopt existing contextual MAB policies one would use the smoothness parameter corresponding to the roughest region. The caveat, however, is that the policy ends up performing more exploration over the smooth regions than what is necessary in these regions. This example demonstrates that, in general, running contextual MAB policies using one single smoothness parameter that captures the worst case smoothness behavior of payoff functions over the entire covariate space is a pessimistic approach and might result in performance loss. Accordingly, in this section, we investigate the possibility of adapting to the local complexity of the problem and regulating the amount of exploration in each region depending on the smoothness behavior of payoff functions in that region. For simplicity of analysis, in this section, we only consider zero-order smoothness assumptions, i.e., $\beta \leq 1$. 


Figure 3: A one-dimensional contextual MAB problem instance in which there are regions over the covariate space where payoff functions behave differently in terms of smoothness.

5.1 A sufficient condition for accounting for local smoothness

We leverage the local self-similarity condition that appears in the literature on locally adaptive non-parametric confidence bands (see Patschkowski and Rohde (2019)) and tailor it to our setting. The definitions and assumptions in this section are mainly inspired by the work of Patschkowski and Rohde (2019). The first definition provides an adjusted notion of Hölder smoothness that will be crucial for the local self-similarity condition in order not to rule out prototypical functions (more details are provided in Patschkowski and Rohde (2019)).

**Definition 5.1** (1-capped Hölder smoothness). The 1-capped Hölder class of functions $\mathcal{H}_{1,U}(\beta, L)$ for the parameters $\beta > 0$ and $L > 0$ in an open cube $U \subseteq (0,1)^d$ is defined as the set of functions $f : U \to \mathbb{R}$ that for any $x, x' \in U$, satisfy the following inequality

$$|f(x') - f(x')| \leq L\|x - x'\|_\infty^\beta.$$  

We drop the indication $U$ whenever $U = (0,1)^d$.

Define the local Hölder exponent of a function $f$ in an open interval $U$ as $\beta_f(U) := \sup\{\beta : f|_U \in \mathcal{H}_{1,U}(\beta)\}$. Note that $\beta_f(U) \in [0,1] \cup \{\infty\}$. To see why this is the case, assume that $f|_U \in \mathcal{H}_{1,U}(\beta, L)$ for some $\beta > 1$, which implies that $f|_U$ has to be constant over $U$, that is, $f|_U \in \mathcal{H}_{1,U}(\infty, L)$. Next, we describe the local self-similarity condition using the notion of 1-capped Hölder smoothness.

**Definition 5.2** (Local self-similarity). For time horizon length $T \in \mathbb{N}$, $0 < \beta < 1$, and $L > 0$, a function $f$ is said to be locally self-similar if $f \in \mathcal{H}_{1,\beta}(L)$ and the following holds true: for any open cube
$U \subseteq (0, 1)^d$ with side-length $2^{-j}$, $j \in \mathbb{N} \cup \{0\}$, there exists some $\beta \in [\underline{\beta}, 1] \cup \{\infty\}$ such that the following conditions are satisfied:

$$f_{|U} \in \mathcal{H}_{1,U}(\beta, L)$$

(5.1)

and

$$\|\Gamma_{j}^{0} f(x; U) - f(x)\|_{\Theta(U, 2^{-j'})} \geq \frac{2^{-\beta j'}}{\log T} \quad \forall j' \in \mathbb{N}, \ j' \geq j + 3.$$  

(5.2)

**Assumption 6.** All the payoff functions $f_k$, $k \in \mathcal{K}$, are locally self-similar.

This condition is a meticulous extension of the global self-similarity condition with the same type of interpretation for any local open cube as opposed to the entire covariate space. Hence, similar to our previous discussion in §3, when this condition holds, one can expect to be able to estimate the smoothness of payoff functions in any local open cube, by examining the estimation bias against its variance in that region.

### 5.2 Overview of the proposed policy

Following [Perchet and Rigollet (2013)](#), we describe the *Locally Adaptive Contextual Bandit (LACB)* policy detailed in Algorithm 2 using the notion of rooted tree. In order to facilitate the description of the policy, we first review some definitions from the graph theory terminology:

- **Tree:** A tree is a graph with no cycles.

- **Root:** The unique node with no predecessor is called the root of the tree.

- **Parent:** The predecessor of a node is called its parent.

- **Child:** Any successor of a node is called its child.

- **Ancestor:** The predecessor of a node together with all the ancestors of the predecessor of a node are called its ancestors. The root node has no ancestors.

- **Leaf:** A node with no successors is called a leaf.

- **Depth:** The number of the ancestors of a node.

Consider a rooted tree $\mathcal{T}^*$ with root $B = (0, 1)^d$ and maximum depth $\bar{l} := \lceil \frac{1}{\underline{\beta} + 2} \log T \rceil$. Every node $B$ in the tree at depth $l \geq 0$ is an open cube in $\mathcal{B}_l$. Furthermore the set of the children of node $B$ is denoted by

$$C(B) := \{B' \in \mathcal{B}_{l(B)+1} : B' \subseteq B\},$$
where $l(B)$ is the depth of node $B$. We also denote by $p(B)$ the parent of $B$.

At every time step $t$, we partition the covariate space $[0, 1]^d$ into 3 disjoint sets $Q_{k,t}$, $k \in K$, and $L_t$. If $X_t \in Q_{k,t}$, we simply pull arm $k$. Each $L_t$, $t = 1, \ldots, T$ consists of the leaves of the subtrees of $T^*$ (we refer to the elements of $L_t$ as live bins). For each live bin $B$ in $L_t$, we first determine the number of samples that we will collect in the bin using a policy similar to the SACB policy in the previous section. After collecting the determined number of samples, the payoff functions are estimated inside the bin. Based on these estimates, we will either replace the bin with its children and repeat the same process or eliminate the bin by assigning it to one of the sets $Q_k$. Next, we will describe each of these steps separately.

5.2.1 Determining the number of samples

Similar to the SACB policy, this step consists of three consecutive steps sampling, estimation, and statistical test that will be repeated until the termination of this step.

Sampling. For each live bin $B$ in $L_t$, we collect samples for both arms for multiple rounds. Define the maximum round index as follows:

$$r(B) := \tilde{r}(T, d, \beta) := \left\lceil 2(l(B) + 3) + \frac{2d}{\beta} + 6 \log \frac{T}{\log T} \right\rceil.$$

At every round $r$, $2^r$ samples is collected for each arm by alternating between them every time a context falls into $B$. Since the goal of collecting samples in each bin is to compute a confidence band for the mean rewards inside the bin, there is no point in collecting samples when the variance falls below the bias. Accordingly, at the end of each round $r$, we run a statistical test to examine variance against bias.

Estimation. Let $\tilde{l}(B) := \left\lceil \frac{l(B) + 3}{2} + \frac{2 \log_d \log T}{2} \right\rceil \vee \left\lceil 2(l(B) + 3) + 2 \log_2 \log T \right\rceil$. For every bin $B$ define the set of mesh points:

$$\mathcal{M}(B) := \left\{ x = \left( \frac{m_1}{2^{l(B)}}, \ldots, \frac{m_d}{2^{l(B)}} \right) : x \in \Theta(B, 2^{-(l(B) + 3)}), m_i = \left\lfloor 2^{\tilde{l}(B)} \right\rfloor \text{ for } i \in [d] \right\} .$$

Define the two bandwidth exponents

$$j_1(B) := l(B) + 3, \quad j_2(B) := l(B) + 3 + \left\lceil \frac{2}{\beta} \log_2 \log T \right\rceil.$$

For every $x \in \mathcal{M}(B)$, we form two separate estimates of the payoff functions for two bandwidth exponents
\( j_1^{(B)} \) and \( j_2^{(B)} \) using local averaging (local polynomial regression of degree zero):

\[
j_k^{(B,r)}(x; j) := \hat{\eta}_{LP}(x; \mathcal{D}_k^{(B,r)}, 2^{-j} 0), \quad j \in \{j_1^{(B)}, j_2^{(B)}\}.
\]

**Statistical test.** At the end of each sampling round \( r \), we check if for any of the mesh points \( x \in \mathcal{M}^{(B)} \), the difference between the estimation using the two bandwidths exponents \( j_1^{(B)} \) and \( j_2^{(B)} \) exceeds a pre-determined threshold. Namely, we check if the following holds true:

\[
\sup_{k \in \mathcal{K}, x \in \mathcal{M}^{(B)}} \left| \hat{f}_k^{(B,r)}(x; j_1^{(B)}) - \hat{f}_k^{(B,r)}(x; j_2^{(B)}) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{r/2}},
\]

where \( \gamma \) is a tuning parameter. Denote by \( r^{(B)}_{last} \) the smallest round for which this inequality holds in the open cube \( B \). If this inequality never happens in \( B \), we simply set \( r_{last}^{(B)} = \bar{r} \). The next proposition states that for each bin \( B \), the number of sampling rounds is of order \( 2l^{(B)} \cdot \min_k \beta f_k(B) \).

**Proposition 5.3.** Under Assumptions 2 and 6, there exist constants \( \underline{C}_r, \bar{C}_r, C_{15}, \) and \( C_{16} \) such that the event

\[
S^{(B)} = \left\{ r_{last}^{(B)} \geq \underline{C}_r + 2(l^{(B)} + 3) \min_k \beta f_k(B) + \left( \frac{2d}{\beta} + 1 \right) \log_2 \log T \quad \text{and} \quad r_{last}^{(B)} \leq \bar{C}_r + 2(l^{(B)} + 3) \min_k \beta f_k(B) + \left( \frac{2d}{\beta} + 5 \right) \log_2 \log T \right\}
\]

holds w.p. at least \( 1 - C_{15} (\log T)^{\frac{2d}{\beta} T^{-2C_{16}}} \), where the constants \( C_{15} \) and \( C_{16} \) depend on only \( \beta, L, \rho, \bar{\rho} \), and \( d \), and \( \underline{C}_r \) and \( \bar{C}_r \) depend on only \( \beta, L, \rho, \) and \( \bar{\rho} \).

### 5.2.2 Estimation and elimination

Let \( \hat{r}_{last}^{(B)} = r_{last}^{(B)} - \left( \frac{2d}{\beta} + 7 \right) \log_2 \log T \). In this step, we collect \( 2^{\hat{r}_{last}^{(B)}} \) samples for each arm by alternating between them every time a context falls into \( B \). Afterwards, we estimate the mean reward difference of the two arms over the bin using the following local average:

\[
\hat{\Delta}^{(B)} := \frac{1}{2^{\hat{r}_{last}^{(B)}}} \left( \sum_{\tau=1}^{2^{\hat{r}_{last}^{(B)}}} Y_1^{(B,r_{last}^{(B)}+1)}(\tau) \right) - Y_2^{(B,r_{last}^{(B)}+1)}(\tau) \).
\]

If this difference is above (below) \( \frac{\gamma (\log T)^{\frac{d}{2}}}{2^{\hat{r}_{last}^{(B)}}} \), we eliminate bin \( B \) and assign it to \( Q_{1,t} \) (\( Q_{2,t} \)). Otherwise, we replace this bin by its children and repeat the same policy.
Proposition 5.4. Under Assumptions 2 and 6, and on the event $S^{(B)}$, there exist constants $C, \bar{C}$ that depend on only $\beta, L, \bar{\rho}$, and $\bar{\rho}$, and $C_{17}$ and $C_{18}$ that depend on only $\beta, L, \bar{\rho}$, $\bar{\rho}$, and $d$ such that the event $S_{\text{last}}^{(B)}$ that is defined below happens w.p. at least $1 - C_{17} \left( \log T \right)^{2d} T^{-\gamma^2 C_{18}}$:

a. If $\sup_{x \in B} f_1(x) - f_2(x) \geq \frac{\gamma C (\log T)^{11/2}}{2^{l(B)} \min_k \beta f_k(B)}$, then, $S_{\text{last}}^{(B)} = \left\{ \hat{\Delta}(B) \geq \frac{\gamma (\log T)^{2}}{2^{l(B)_{\text{last}}/2}} \right\}$.

b. If $\inf_{x \in B} f_1(x) - f_2(x) \leq \frac{-\gamma C (\log T)^{11/2}}{2^{l(B)} \min_k \beta f_k(B)}$, then, $S_{\text{last}}^{(B)} = \left\{ \hat{\Delta}(B) \leq \frac{-\gamma (\log T)^{2}}{2^{l(B)_{\text{last}}/2}} \right\}$.

c. If $\sup_{x \in B} |f_1(x) - f_2(x)| < \frac{\gamma C (\log T)^{7/2}}{2^{l(B)} \min_k \beta f_k(B)}$, then, $S_{\text{last}}^{(B)} = \left\{ \left| \hat{\Delta}(B) \right| < \frac{\gamma (\log T)^{2}}{2^{l(B)_{\text{last}}/2}} \right\}$.

Based on this proposition, if the arm rewards are separated enough inside a bin $B$ then this bin will be eliminated w.h.p; otherwise, since we are not sure which one of the arms are optimal inside $B$ we will replace this bin with its children.

5.2.3 Traveled depth

In this subsection, we discuss how deep in the tree $T^*$, the LACB policy travels. Define $T^*$ to be a subtree of $T^*$ with the set of leaves $\tilde{L}^*$ such that:

1. For every leaf node (bin) $B \in \tilde{L}^*$, $T \cdot 2^{-l(B)}d \leq 2C_r \left( \log T \right)^{2d + 5} 2^{l(B) \min_k \beta f_k(B)}$

2. For every non-leaf node $B \in \tilde{T}^* \setminus \tilde{L}^*$, $T \cdot 2^{-l(B)}d > 2C_r \left( \log T \right)^{2d + 5} 2^{l(B) \min_k \beta f_k(B)}$.

Proposition 5.5. On the event $\bigcap_{B \in \tilde{T}^*} S^{(B)}$, the probability that any bin in the set $\left\{ B : p(B) \in \tilde{L}^* \right\}$, where $p(B)$ is the parent of $B$, becomes live is less than $T^{-d/8} \exp \left( -T^{-d/8} / 5 \right)$.

Note that the depth of the subtree $T^*$ depends on how smooth the payoff functions behave locally. That is, over the smooth regions, the traveled depth is shallower. This is also consistent with intuition, since in these regions more extrapolation can be done, and hence, there is less need for partitioning.
The pseudo-code that describes the policy is presented below.

**Algorithm 2: Locally Adaptive Contextual Bandits (LACB)**

1. **Input:** Time horizon $T$, minimum smoothness exponents $\beta$, and a tuning parameter $\gamma$
2. **Initialization:** For $B \in \{0, 1\}$ initialize $N_k^{(B)} \leftarrow 0 \ \forall k \in K$, and set $l^{(B)} \leftarrow 0$, $\xi^{(B)} \leftarrow 0$, and $Q_{k,1} \leftarrow \emptyset \ \forall k \in K$
3. for $t = 1, \ldots, T$ do
   
   if $X_t \in \bigcup_{k \in K} Q_{k,t}$ then
   
   Pull the arm corresponding to $Q_{k,t}$ that contains $X_t$: $\pi_t = k$ iff $X_t \in Q_{k,t}$
   
   continue

   Determine the bin in which the current covariate is located: $B \in \mathcal{L}_t(X_t)$ s.t. $X_t \in B$

   Alternate between the arms: $\pi_t \leftarrow 1 + 1 \ \{N_1^{(B)} > N_2^{(B)}\}$

   Update the counters: $N_k^{(B)} \leftarrow N_k^{(B)} + 1 \ \{\pi_t = k\} \ \forall k \in K$

   Receive and observe reward $Y_{\pi_t,t}$

   if $\left[N_1^{(B)} + N_2^{(B)} \geq 2 \times 2^r^{(B)} \text{ or } r^{(B)} \geq r^{(B)} \right]$ and $\xi^{(B)} = 0$ then

   if $\sup_{k \in K, x \in \mathcal{M}^{(B)}} \left| f_k^{(B,r^{(B)})}(x; \hat{j}_1(B)) - f_k^{(B,r^{(B)})}(x; \hat{j}_2(B)) \right| \leq \frac{3\gamma T^{2d+2}}{2 \cdot r^{(B)/2}}$ for all $k \in K$ then

   Record $r^{(B)}_{last}, r^{(B)}_{last} \leftarrow r^{(B)}$, $\xi^{(B)} \leftarrow 1$

   Collect double number of samples in the next round: $r^{(B)} \leftarrow r^{(B)} + 1$; Reset the counters:

   $N_k^{(B)} \leftarrow 0 \ \forall k \in K$

   else if $N_1^{(B)} + N_2^{(B)} \geq 2 \times 2^r^{(B)} - 1$ and $\xi^{(B)} = 1$ then

   Estimate the mean reward difference: $\hat{\Delta}^{(B)} \leftarrow \frac{1}{2^r_{last}} \left( \sum_{\tau=1}^{2^r_{last}} Y_{1,\tau}^{(B,r^{(B)}_{last}+1)} - Y_{2,\tau}^{(B,r^{(B)}_{last}+1)} \right)$

   if $\hat{\Delta}^{(B)} \geq \frac{\gamma T^{2d+2}}{2^r_{last}/2}$ then

   $Q_{1,t+1} \leftarrow Q_{1,t} \cup B$, $Q_{2,t+1} \leftarrow Q_{2,t}$

   else if $\hat{\Delta}^{(B)} \leq -\frac{\gamma T^{2d+2}}{2^r_{last}/2}$ then

   $Q_{1,t+1} \leftarrow Q_{1,t}$, $Q_{2,t+1} \leftarrow Q_{2,t} \cup B$

   else

   $Q_{1,t+1} \leftarrow Q_{1,t}$, $Q_{2,t+1} \leftarrow Q_{2,t}$

   for $B' \in C(B)$ do

   $N_k^{(B')} \leftarrow 0 \ \forall k \in K$, and set $l^{(B')} \leftarrow 0$, $\xi^{(B')} \leftarrow 0$

   else

   Pass the regions $Q_{k,t}$ to next round: $Q_{k,t+1} \leftarrow Q_{k,t} \ \forall k \in K$

The next theorem characterizes the regret rate of the LACB policy.
Theorem 5.6. If the LACB policy is run with large enough $\gamma \geq \gamma_0(\beta, \bar{\beta}, L, \rho, \bar{\rho}, d)$ then, under Assumptions 1, 2, and 6, it has an expected regret at time $T$ bounded by

$$R^\pi(T) \leq C (\log T)^{\frac{10d + 1}{2}} \sum_{B \in \mathcal{T}^*} 2^{l(B)} (2 \min_k \beta f_k(B) - \min_k \beta f_k(p(B))),$$

for some constant $C$ independent of $T$ and large enough $T \geq T_0(\bar{\beta}, b, L)$. 

28
A Proofs of main results

A.1 Proof of Lemma 4.1

Fix some \( x \in U \). Let \( \hat{\theta}(u; p, l, U) := \sum_{|s| \leq p} \xi_s u^s \) be a polynomial of degree \( p \) on \( \mathbb{R}^d \) that minimizes

\[
\int_U \left| f(u) - \hat{\theta} \left( \frac{u - x}{h}; p, l, U \right) \right|^2 K \left( \frac{u - x}{h} \right) p_X(u \mid U) du = \int_U f^2(u) K \left( \frac{u - x}{h} \right) p_X(u \mid U) du
\]

\[+ \sum_{|s_1|, |s_2| \leq p} \xi_{s_1} \xi_{s_2} \int_U \left( \frac{u - x}{h} \right)^{s_1 + s_2} K \left( \frac{u - x}{h} \right) p_X(u \mid U) du
\]

\[- 2 \sum_{|s| \leq p} \xi_s \int_U f(u) \left( \frac{u - x}{h} \right)^s K \left( \frac{u - x}{h} \right) p_X(u \mid U) du,
\]

where \( h = 2^{-l} \). Equivalently, \( \hat{\theta}(u; p, l, U) \) can be characterized by its vector of coefficients \( \xi \) that minimizes

\[
\sum_{|s_1|, |s_2| \leq p} \xi_{s_1} \xi_{s_2} \int_{\mathbb{R}^d} u^{s_1 + s_2} K(u)p_X(x + hu \mid U) du - 2 \sum_{|s| \leq p} \xi_s \int_{\mathbb{R}^d} f(u) u^s K(u)p_X(x + hu \mid U) du = \xi^T B \xi - 2W^T \xi,
\]

where we define the matrix \( B := (B_{s_1,s_2})_{|s_1|, |s_2| \leq p} \) and the vector \( W := (W_s)_{|s| \leq p} \) with elements

\[
B_{s_1,s_2} := \int_{\mathbb{R}^d} u^{s_1 + s_2} K(u)p_X(x + hu \mid U) du, \quad W_s := \int_{\mathbb{R}^d} f(u) u^s K(u)p_X(x + hu \mid U) du.
\]

Note that if \( B \) is a positive definite matrix then, the minimizer of \((A.1)\) is \( \xi = B^{-1}W \), which implies the desired result: \( \nabla_B^2 f(x; U) = R^T (0) B^{-1} W \). In order to show that this is indeed the case, we note that

\[
\lambda_{\min}(B) = \min_{\|Z\|=1} Z^T B Z = \int_U \left( \sum_{|s| \leq p} Z_s u^s \right)^2 K(u)p_X(x + hu \mid U) du \geq \frac{\rho^{2d'}}{\tilde{\rho}} \int_A \left( \sum_{|s| \leq p} Z_s u^s \right)^2 \ du,
\]

where \( A = \{ u \in \mathbb{R}^d : \|u\|_\infty \leq 1; x + hu \in U \} \). Note that

\[
\lambda[A] \geq h^{-d} \lambda [\Xi(x, h) \cap U] \geq 2^{-d} h^{-d} \lambda [\Xi(x, h)] = 2^{-d} \lambda [\Xi(0, 1)]
\]

. Let \( A \) denote the class of compact subsets of \( \Xi(0, 1) \) having the Lebesgue measure \( 2^{-d} \lambda [\Xi(0, 1)] \). Using the previous display, we obtain

\[
\lambda_{\min}(B) \geq \frac{\rho^{2d'}}{\tilde{\rho}} \min_{\|Z\|\leq 1; S \in A} \int_S \left( \sum_{|s| \leq p} Z_s u^s \right)^2 du =: \frac{\rho^{2d'}}{\tilde{\rho}} \tilde{\mu}_0.
\]

\[\text{(A.2)}\]
By the compactness argument, the minimum in the above expression exists, and is strictly positive.

In order to prove the last claim in the lemma, note that for any \(\hat{x} \in U\),

\[
|\Gamma_p^t f(x; U) - \Gamma_p^t f(\hat{x}; U)| = \left| \hat{\theta}(0; p, l, U) - \hat{\theta}\left(\frac{\hat{x} - x}{h}; p, l, U\right) \right|
\]

\[
= \left| \sum_{|s| \leq p, s \neq (0, \ldots, 0)} \xi_s \left(\frac{\hat{x} - x}{h}\right)^s \right| \leq M h^{-1} \|\hat{x} - x\|_\infty \|\xi\|.
\]

Also, by (A.2), we have

\[
\|\xi\| \leq \left\| B^{-1} W \right\| \leq \frac{2^{-d_{\alpha'}}}{\rho} \tilde{\rho}_0^{-1} M^{1/2} \max_s |W_s|,
\]

and

\[
|W_s| = \left| \int_{\mathbb{R}^d} u^s f(x + hu) K(u) p_X(x + hu | U) du \right| \leq \int_{\mathbb{R}^d} K(u) p_X(x + hu | U) du \leq 2^d.
\]

Putting together the above three displays, one obtains

\[
|\Gamma_p^t f(x; U) - \Gamma_p^t f(\hat{x}; U)| \leq \tilde{\rho} \tilde{\rho}_0^{-1} M^{3/2} h^{-1} \|\hat{x} - x\|_\infty.
\]

To prove the last part, define the vector \(Z := (Z_s)_{|s| \leq p}\) with elements

\[
Z_s := \frac{h^{|s|} f^{(s)}(x)}{s!} \cdot 1 \{|s| \leq |\beta|\}.
\]

Note that

\[
f(x) = R^\top (0) B^{-1} B Z.
\]

As a result, we have

\[
|f(x) - \Gamma_p^t f(x; U)| = \left| R^\top (0) B^{-1} (BZ - W) \right| \leq \|B^{-1}\| \cdot \|BZ - W\| \leq \frac{2^{-d_{\alpha'}}}{\rho} \tilde{\rho}_0^{-1} M^{1/2} \max_s |(BZ)_s - W_s|,
\]

where the last inequality follows from (A.2). Furthermore, one has

\[
|(BZ)_s - W_s| = \left| \int_{\mathbb{R}^d} u^s \left( \sum_{|s'| \leq |\beta|} \frac{(hu)^s f^{(s)}(x)}{s!} - f(x + hu) \right) K(u) p_X(x + hu | U) du \right|
\]

\[
\leq \int_{\mathbb{R}^d} |u^s| \left| \sum_{|s'| \leq |\beta|} \frac{(hu)^s f^{(s)}(x)}{s!} - f(x + hu) \right| K(u) p_X(x + hu | U) du
\]

\[
\leq \int_{\mathbb{R}^d} L h^\beta p_X(x + hu | U) du = L h^\beta 2^{d_{\alpha'}}.
\]
where the last inequality follows from the assumption that $f \in H(\beta, L)$. Putting the last two displays together, the result follows. This concludes the proof. 

\[ \square \]

### A.2 Proof of Proposition 4.5

The proof is a simple extension of the proof of Theorem 3.2 in [Audibert and Tsybakov (2007)](https://example.com); however, we provide the proof for completeness. Fix $x \in \mathcal{X}$ and $\delta > 0$. Consider the matrices $B := (B_{s_1, s_2})_{|s_1|,|s_2| \leq p}$ and $\bar{B} := (\bar{B}_{s_1, s_2})_{|s_1|,|s_2| \leq p}$ with the elements

$$B_{s_1, s_2} := \int_{\mathbb{R}^d} u^{s_1+s_2} K(u) \mu(x + hu) du, \quad \bar{B}_{s_1, s_2} := \frac{1}{nh^d} \sum_{i=1}^n \left( \frac{X_i - x}{h} \right)^{s_1+s_2} K \left( \frac{X_i - x}{h} \right).$$

The smallest eigenvalue of $\bar{B}$ satisfies

$$\lambda_{\min}(\bar{B}) = \min_{\|W\| = 1} W^T \bar{B} W \quad \geq \min_{\|W\| = 1} W^T B W + \min_{\|W\| = 1} W^T (\bar{B} - B) W \quad \geq \min_{\|W\| = 1} W^T B W - \sum_{|s_1|,|s_2| \leq p} |\bar{B}_{s_1, s_2} - B_{s_1, s_2}|. \quad (A.3)$$

Define $\mathcal{X}_n := \{ u \in \mathbb{R}^d : \|u\| \leq 1; x + hu \in \mathcal{X} \}$. For any vector $W$ satisfying $\|W\| = 1$, we obtain

$$W^T B W = \int_{\mathbb{R}^d} \left( \sum_{|s| \leq p} W_s u_s \right)^2 K(u) \mu(x + hu) du \geq \mu \int_{\mathcal{X}_n} \left( \sum_{|s| \leq p} W_s u_s \right)^2 du.$$

By assumption of the proposition, $h \leq r_0$. Since $\mathcal{X}$ is a closed cube we get

$$\lambda(\mathcal{X}_n) \geq h^{-d} \lambda[\text{Ball}_2(x, h) \cap \mathcal{X}] \geq 2^{-d} h^{-d} \lambda[\text{Ball}_2(x, h)] \geq 2^{-d} \lambda[\text{Ball}_2(0, 1)],$$

where $\text{Ball}_2(x, h)$ is the Euclidean ball of radius $h$ centered at $x$.

Let $\mathcal{A}$ denote the class of all compact subsets of $\text{Ball}_2(0, 1)$ having the Lebesgue measure $c_0 \lambda[\text{Ball}_2(0, 1)]$. Using the previous display, we obtain

$$\min_{\|W\| = 1} W^T B W \geq \mu \min_{\|W\| = 1; S \in \mathcal{A}} \int_S \left( \sum_{|s| \leq p} W_s u_s \right)^2 du =: 2c_\mu \quad (A.4)$$

By the compactness argument, the above minimum exists and is strictly positive.
For $i = 1, \ldots, n$ and any multi-indices $s_1, s_2$ such that $|s_1|, |s_2| \leq p$, define

\[
T_i^{(s_1, s_2)} := \frac{1}{h^d} \left( \frac{X_i - x}{h} \right)^{s_1 + s_2} K \left( \frac{X_i - x}{h} \right) - \int_{\mathbb{R}^d} u^{s_1 + s_2} K(u) \mu(x + hu) du.
\]

We have $\mathbb{E}T_i^{(s_1, s_2)} = 0$, $|T_i^{(s_1, s_2)}| \leq 2h^{-d}$, and the following bound on the variance of $T_i^{(s_1, s_2)}$:

\[
\text{Var} T_i^{(s_1, s_2)} \leq \frac{1}{h^d} \mathbb{E} \left[ \left( \frac{X_i - x}{h} \right)^{2s_1+2s_2} K^2 \left( \frac{X_i - x}{h} \right) \right] \\
\leq \frac{1}{h^d} \int_{\mathbb{R}^d} u^{2s_1+2s_2} K^2(u) \mu(x + hu|B) du \\
\leq \frac{\bar{\mu}}{h^d} \max_{j \leq p} \int_{\mathbb{R}^d} (1 + |u^j|) K^2(u) du =: \frac{\kappa \bar{\mu}}{h^d}.
\]

From Bernstein’s inequality, we get

\[
\mathbb{P}\{|\hat{B}_{s_1, s_2} - B_{s_1, s_2}| > \epsilon\} = \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^n T_i^{(s_1, s_2)} \right| > \epsilon \right\} \leq 2 \exp \left( -\frac{nh^d \epsilon^2}{2 \kappa \bar{\mu} + 4 \epsilon / 3} \right).
\]

This inequality along with (A.3) and (A.4) imply that

\[
\mathbb{P}\{\lambda_{\min}(\hat{B}) \leq c \mu\} \leq 2M^2 \exp \left( -\frac{nh^d M^{-4} \epsilon^2 \mu^2}{2 \kappa \bar{\mu} + 4 M^{-2} c \mu / 3} \right), \quad (A.5)
\]

where $M^2$ is the number of elements in the matrix $\hat{B}$. In what follows assume that $\lambda_{\min}(\hat{B}) \geq c \mu$.

Therefore,

\[
\mathbb{P}\{|\hat{\eta}^{LP}(x; D, h, p) - \eta(x)| \geq \delta\} \leq \mathbb{P}\{\lambda_{\min}(\hat{B}) \leq c \mu\} + \mathbb{P}\{\lambda_{\min}(\hat{B}) > c \mu\}, \quad (A.6)
\]

We now evaluate the second term on the right hand side of the above inequality. Define the matrix $Z := (Z_{i,s})_1 \leq i \leq n, |s| \leq p$ with elements

\[
Z_{i,s} := (X_i - x)^s \sqrt{K \left( \frac{X_i - x}{h} \right)}.
\]

The $s$-th column of $Z$ is denoted by $Z_s$, and we introduce $Z^{(\eta)} := \sum_{|s| \leq |\beta|} \frac{\eta^{(s)}(x)}{s!} Z_s$. Since $Q = Z^\top Z$ we get

\[
\forall |s| \leq |\beta| : U^\top (0) Q^{-1} Z^\top Z = 1 \{s = (0, \ldots, 0)\},
\]
hence \( R^T(0)Q^{-1}Z^T Z^{(q)} = \eta(x) \). So we can write

\[
\hat{\eta}^{LP}(x; D, h, p) - \eta(x) = R^T(0)Q^{-1} \left( V - Z^T Z^{(q)} \right) = R^T(0)\bar{B}^{-1}\mathbf{a},
\]

where \( \mathbf{a} := \frac{1}{nh^d} H \left( V - Z^T Z^{(q)} \right) \in \mathbb{R}^M \) and \( H \) is a diagonal matrix \( H := (H_{s_1, s_2})_{|s_1||s_2| \leq p} \) with elements \( H_{s_1, s_2} := h^{-s_1} \mathbb{1} \{ s_1 = s_2 \} \). For \( \lambda_{\min}(\bar{B}) > c\mu \), we get

\[
|\hat{\eta}^{LP}(x; D, h, p) - \eta(x)| \leq \|\bar{B}^{-1}\mathbf{a}\| \leq \lambda_{\min}^{-1}(\bar{B})\|\mathbf{a}\| \leq c^{-1} \mu^{-1} M \max_s \|a_s\|, \tag{A.7}
\]

where \( a_s \) are the components of the vector \( \mathbf{a} \) given by

\[
a_s = \frac{1}{nh^d} \sum_{i=1}^n [Y_i - \eta_x(X_i)] \left( \frac{X_i - x}{h} \right)^s K \left( \frac{X_i - x}{h} \right).
\]

Note that \( \eta_x(X_i) \) is the Taylor expansion of \( \eta \) at \( x \) and of degree \( |\beta| \) (not necessarily \( p \)) evaluated at \( X_i \).

Define

\[
T_i^{(s,1)} := [Y_i - \eta_x(X_i)] \left( \frac{X_i - x}{h} \right)^s K \left( \frac{X_i - x}{h} \right),
\]

\[
T_i^{(s,2)} := [\eta(X - i) - \eta_x(X_i)] \left( \frac{X_i - x}{h} \right)^s K \left( \frac{X_i - x}{h} \right).
\]

One has

\[
|a_s| \leq \left| \frac{1}{n} \sum_{i=1}^n T_i^{(s,1)} \right| + \left| \frac{1}{n} \sum_{i=1}^n [T_i^{(s,2)} - \mathbb{E}T_i^{(s,2)}] \right| + \left| \mathbb{E}T_i^{(s,2)} \right|. \tag{A.8}
\]

Note that \( \mathbb{E}T_i^{(s,1)} = 0 \), \( |T_i^{(s,1)}| \leq 2h^{-d} \), and

\[
\mathbb{V} \text{ar} T_i^{(s,1)} \leq \frac{1}{4h^d} \int_{\mathbb{R}^d} u^{2s} K^2(u) \mu(x + hu) du \leq \frac{\kappa \mu}{4h^d},
\]

\[
|T_i^{(s,2)} - \mathbb{E}T_i^{(s,2)}| \leq Lh^{\beta-d} + L\kappa h^{\beta} \leq Ch^{\beta-d},
\]

\[
\mathbb{V} \text{ar} T_i^{(s,2)} \leq L^2 h^{2\beta-d} \int_{\mathbb{R}^d} |u^{2s}| K^2(u) \mu(x + hu) du \leq L^2 \mu \kappa h^{2\beta-d}.
\]

From Bernstein’s inequality, for \( \epsilon_1, \epsilon_2 > 0 \), we obtain

\[
\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n T_i^{(s,1)} \right| \geq \epsilon_1 \right\} \leq 2 \exp \left( \frac{-nh^d \epsilon_1^2}{\kappa \mu / 2 + 4 \epsilon_1 / 3} \right).
\]
and
\[
P\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} T_i^{(s,2)} - E T_i^{(s,2)} \right| \geq \epsilon_2 \right\} \leq 2 \exp \left( \frac{-n h^d \epsilon_2^2}{2 L^2 k \bar{\mu} h^3 + 2 Ch^3 \epsilon_2 / 3} \right).
\]

Since also
\[
\left| E T_i^{(s,2)} \right| \leq L h^\beta \int_{\mathbb{R}^d} |u^s| K^2(u) \mu(x + hu) du \leq L k \bar{\mu} h^\beta
\]
we get, using (A.8), that if \(3 L k \bar{\mu} h^\beta c^{-1} M \leq \delta \leq 1\) the following inequality holds
\[
P \left\{ \left| a_s \right| \geq \frac{c \mu \delta}{M} \right\} \leq P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} T_i^{(s,1)} \right| > \frac{c \mu \delta}{3M} \right\} + P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \left[ T_i^{(s,2)} - E T_i^{(s,2)} \right] \right| > \frac{c \mu \delta}{3M} \right\}
\]
\[
\leq 4 \exp \left( -C n h^d \mu^2 \bar{\mu}^{-1} \delta^2 \right).
\]

Combining this inequality with (A.5), (A.6), and (A.7), we get
\[
P \left\{ \left| \eta^{LP}(x; D, h, p) - \eta(x) \right| \geq \delta \right\} \leq C_1 \exp \left( -C_2 n h^d \mu^2 \bar{\mu}^{-1} \delta^2 \right)
\]
for \(3 L k \bar{\mu} h^\beta c^{-1} M \leq \delta \) (for \(\delta > 1\), this inequality is obvious since \(\eta, \eta^{LP}\) take values in \([0, 1]\)). The constants \(C_1, C_2\) do not depend on the density \(\mu\), on its support \(\mathcal{X}\) and the point \(x \in \mathcal{X}\). This concludes the proof.

\[\Box\]

### A.3 Proof of Proposition 4.6

Fix an open cube \(U \subseteq (0, 1)^d\) with side-length \(2^{-l'}\), \(l' \in \mathbb{R}_+\). Consider the matrix \(B := (B_{s_1,s_2})_{|s_1|,|s_2| \leq p}\) and the vector \(W := (W_s)_{|s| \leq p}\) with elements
\[
B_{s_1,s_2} := \int_{\mathbb{R}^d} u^{s_1+s_2} K(u) \mu(x + hu) du, \quad W_s := \int_{\mathbb{R}^d} u^s \eta(x + hu) K(u) \mu(x + hu) du,
\]
as well as the matrix \(\bar{B} := (\bar{B}_{s_1,s_2})_{|s_1|,|s_2| \leq p}\) and the vector \(\bar{W} := (\bar{W}_s)_{|s| \leq p}\) with elements
\[
\bar{B}_{s_1,s_2} := \frac{1}{nh^d} \sum_{i=1}^{n} \left( \frac{X_i - x}{h} \right)^{s_1+s_2} K \left( \frac{X_i - x}{h} \right), \quad \bar{W}_s := \frac{1}{nh^d} \sum_{i=1}^{n} Y_i \left( \frac{X_i - x}{h} \right)^s K \left( \frac{X_i - x}{h} \right).
\]
By Lemmas 4.1 and 4.4, we have

\[
|\hat{\eta}^{LP}(x; D, h, p) - \Gamma_i^p \eta(x)| = \left| U(0)^\top Q^{-1}V - U(0)^\top B^{-1}W \right|
= \left| U(0)^\top B^{-1} \bar{W} - U(0)^\top B^{-1}W \right|
\leq \left| U(0)^\top B^{-1}(\bar{W} - W) \right| + \left| U(0)^\top (B^{-1} - B^{-1}) \bar{W} \right| =: J_1 + J_2.
\]

That is,

\[
P \{ |\hat{\eta}^{LP}(x; D, h, p) - \Gamma_i^p \eta(x; U)| \geq \delta \} \leq P \{ J_1 \geq 3\delta/4 \} + P \{ J_2 \geq \delta/4 \}.
\]

(A.9)

First, we analyze \( J_1 \). Note that

\[
J_1 \leq \| B^{-1}(\bar{W} - W) \| \leq \lambda_{\min}^{-1}(B) \| \bar{W} - W \| \leq \mu_0^{-1}2^{-d\nu} \| \bar{W} - W \| \leq \mu_0^{-1}2^{-d\nu} M \max_{s} |\bar{W}_s - W_s|,
\]

(A.10)

where the third inequality follows from \( \lambda_{\min}(B) \geq \mu_0 2^{d\nu} \) by Lemma 4.1, and \( M \) is the number of elements in the vector \( W \). Define

\[
T_i^{(s)} := \frac{1}{h^d} Y_i \left( \frac{X_i - x}{h} \right)^s K \left( \frac{X_i - x}{h} \right) - \int_{\mathbb{R}^d} \eta(x + hu)u^s K(u)p_X(x + hu|U) du.
\]

We have \( \mathbb{E} \left[ T_i^{(s)} \right] = 0 \), \( |T_i^{(s)}| \leq 2h^{-d} \), and

\[
\text{Var} \left[ T_i^{(s)} \right] \leq \frac{1}{h^{2d}} \mathbb{E} \left[ \left( \frac{X_i - x}{h} \right)^{2s} K^2 \left( \frac{X_i - x}{h} \right) \right] \leq \frac{1}{h^{2d}} \int_{\mathbb{R}^d} u^{2s} K^2(u)p_X(x + hu|U) du \leq \frac{2^{d\nu}}{h^d}.
\]

By Bernstein’s inequality, we get

\[
P \left\{ |W_s - W_s| \geq \epsilon \right\} = P \left\{ \left| \frac{1}{n} \sum_{i=1}^n T_i^{(s)} \right| > \epsilon \right\} \leq 2 \exp \left( \frac{-n\epsilon^2}{2^{1+d\nu} + 4\epsilon/3} \right).
\]

Combining this inequality with (A.10), one obtains

\[
P \{ J_1 \geq 3\delta/4 \} \leq \sum_{|s| \leq \rho} P \left\{ |W_s - W_s| \geq 3\mu_0 2^{d\nu} M^{-1} \delta/4 \right\}
= \sum_{|s| \leq \rho} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n T_i^{(s)} \right| \geq 3\mu_0 2^{d\nu} M^{-1} \delta/4 \right\} \leq 2M \exp \left( \frac{-9\mu_0^2 M^{-2}2^{-d\nu} \rho \delta^2/16}{2 + \mu_0 M^{-1} \delta} \right).
\]

(A.11)
Now, we analyze $J_2$. Note that

$$J_2 \leq \| (\bar{B}^{-1} - B^{-1}) \bar{W} \| \leq \| \bar{B}^{-1} - B^{-1} \| \| \bar{W} \| \leq M \| \bar{B}^{-1} - B^{-1} \| \cdot \max_s |\bar{W}_s| \leq M \| \bar{B}^{-1} - B^{-1} \| h^{-d}. \quad \text{(A.12)}$$

Define $Z := \bar{B} - B$. One has

$$\lambda_{\max}(Z) \leq \sum_{|s_1|, |s_2| \leq p} |Z_{s_1, s_2}|.$$

Define

$$T_i^{(s_1, s_2)} := \frac{1}{h^d} \left( \frac{X_i - x}{h} \right)^{s_1 + s_2} K \left( \frac{X_i - x}{h} \right) - \int_{\mathbb{R}^d} u^{s_1 + s_2} K(u) p_X(x + hu | U) du.$$

We have $\mathbb{E} \left[ T_i^{(s_1, s_2)} \right] = 0$, $|T_i^{(s_1, s_2)}| \leq 2h^{-d}$, and

$$\text{Var} \left[ T_i^{(s_1, s_2)} \right] \leq \mathbb{E} \left[ \frac{1}{h^d} \left( \frac{X_i - x}{h} \right)^{2s_1 + 2s_2} K^2 \left( \frac{X_i - x}{h} \right) \right] = \frac{1}{h^d} \int_{\mathbb{R}^d} u^{2s_1 + 2s_2} K^2(u) p_X(x + hu | U) du \leq \frac{2d!}{h^d}.$$

By Bernstein’s inequality, one obtains

$$\mathbb{P} \left\{ \lambda_{\max}(Z) \geq 2^d M^{-1} \mu_0^2 \delta / 8 \right\} \leq \mathbb{P} \left\{ \sum_{|s_1|, |s_2| \leq p} |Z_{s_1, s_2}| \geq 2^d M^{-1} \mu_0^2 \delta / 8 \right\} \leq \sum_{|s_1|, |s_2| \leq p} \mathbb{P} \left\{ |Z_{s_1, s_2}| \geq 2^d M^{-3} \mu_0^2 \delta / 8 \right\} = \sum_{|s_1|, |s_2| \leq p} \mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n T_i^{(s_1, s_2)} \right| \geq h^{-d} M^{-3} \mu_0^2 \delta / 8 \right\} \leq 2M^2 \exp \left( -n2^d(h^{-d} M^{-3} \mu_0^2 \delta^2 / 64) / (2 + M^{-3} \mu_0^2 \delta^2 / 64) \right).$$

By Lemma \[4.1\], $\|B^{-1}\| \leq 2^{-d'} \mu_0^{-1}$. That is, on the event $\{ \lambda_{\max}(Z) \leq 2^d M^{-1} \mu_0^2 \delta / 8 \}$, we have $\|B^{-1/2} Z B^{-1/2}\| \leq M^{-1} \mu_0 \delta / 8$ in which case if $M^{-1} \mu_0 \delta / 8 < 1/2$, one obtains

$$\| \bar{B}^{-1} - B^{-1} \| = \| B^{-1/2} \left( (I + B^{-1/2} Z B^{-1/2})^{-1} - I \right) B^{-1/2} \| \leq \| B^{-1} \| \left\| (I + B^{-1/2} Z B^{-1/2})^{-1} - I \right\| \leq 2^{-d'} \mu_0^{-1} \sum_{j=1}^{\infty} \| B^{-1/2} Z B^{-1/2} \| j \leq 2^{-d'} \mu_0^{-1} \sum_{j=1}^{\infty} (M^{-1} \mu_0 \delta / 8)^j \leq 2^{-d'} M^{-1} \delta / 4.$$
This inequality along with (A.12) imply \( J_2 \leq \delta/4 \). In other words,
\[
P\{J_2 \geq \delta/4\} \leq P\left\{\lambda_{\text{max}}(Z) \geq 2^{d/2} M^{-1} \mu_0^2 \delta/8\right\} \leq 2M^2 \exp\left(-\frac{-n2^{d(l'-l)} M^{-6} \mu_0^4 \delta^2/64}{2 + M^{-3} \mu_0^2 \delta/6}\right).
\]
Combining this inequality with (A.9) and (A.11) gives
\[
P\{\left|\hat{\eta}^{\text{LP}}(x; D, h, p) - \Gamma p_{\ell}(x; U)\right| \geq \delta\} \leq C_4 \exp\left(-C_5 n2^{d(l'-l)} \delta^2\right)
\]
if \( M^{-1} \mu_0 \delta/8 < 1/2 \). This concludes the proof.

A.4 Proof of Theorem 3.1

Step 1 (Preliminaries). Fix time horizon length \( T \geq 1 \), and two Hölder exponents \( 0 < \gamma \leq \frac{1}{2} \), \( 0 < \beta < \frac{\gamma}{2 - \alpha \gamma} \wedge \frac{\alpha^2}{1 - 4\delta^2} \), some parameter \( 0 \leq \alpha \leq \frac{1}{7} \), and some positive constants \( L, C_0, \rho, \bar{\rho} \). For any policy \( \pi \) and time horizon \( T \), let \( S_{\pi}(P; T) \) be the inferior sampling rate defined as
\[
S_{\pi}(P; T) := \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} I\{f_{\pi_t}(X_t) \neq f_{\pi_t}(X_t)\}\right].
\]
Fix a covariate distribution \( P_X \) that satisfies Assumption 1 with parameters \( \rho, \bar{\rho} \). For any policy \( \pi \) and function \( f : [0, 1] \to [0, 1] \), denote by \( S_{\pi}(f; T) \) the inferior sampling rate of \( \pi \) when covariates’ distribution is \( P_X \), \( \mathbb{E}[Y_{1,t} | X_t] = f(X_t) \), and \( \mathbb{E}[Y_{2,t} | X_t] = 1/2 \). Clearly, the oracle policy \( \pi^*_f \) is given by \( \pi^*_f(x) = 2 - I\{f(x) \geq 1/2\} \). Furthermore, we denote by \( P_{\pi,f} \) and \( \mathbb{E}_{\pi,f} \) the corresponding probability and expectation. Finally, let \( R_{\pi,\gamma,\alpha}(T) := \sup \{ R_{\pi}(P; T) : P \in \mathcal{P}(\gamma, \alpha, 1) \} \)

Step 2 (From regret to inferior sampling rate). In view of the following lemma, it will be sufficient to first analyze inferior sampling rate and then, revert the result back to regret at the end.

Lemma A.1 ([Rigollet and Zeevi 2010, Lemma 3.1]). For any \( \alpha > 0 \) under the margin condition in Assumption 3, we have
\[
S_{\pi}(P; T) \leq C_{sr} T^{1/\alpha \gamma} [R_{\pi}(P; T)]^{\alpha \gamma},
\]
for any policy \( \pi \) and some positive constant \( C_{sr} \).

Note that if for any problem instance in \( P \in \mathcal{P}(\gamma, \alpha, 1) \), the policy \( \pi \) satisfies: \( R_{\pi}(P; T) \leq R_{\pi,\gamma,\alpha}(T) \) then, by Lemma A.1 it has to also satisfy \( S_{\pi}(P; T) \leq C_{sr} T^{1/\alpha \gamma} [R_{\pi,\gamma,\alpha}(T)]^{\alpha \gamma} = S_{\pi,\gamma,\alpha}(T) \).
Step 3 (Constructing the problem instances). We will reduce our problem to a hypothesis testing problem. To do so, we will construct some problem instances first. Define the parameter $\Delta > 0$ such that

\[
\frac{64\tilde{C}_0^2 \Delta^2 S^\gamma (T)}{3M} = \frac{1}{2},
\]

where we define $M := \lfloor \Delta^{\frac{\beta - \frac{\gamma}{2}}{2}} \rfloor \lor 1$ and $\tilde{C}_0 := \left( \frac{C_0}{2^\gamma} \right)^{\frac{1}{\gamma}} \land \left( \frac{C_0}{6^\gamma} \right)^{\frac{1}{2 \gamma}} \land \left( \frac{C_0}{12^\gamma} \right)^{\frac{1}{2 \gamma}} \land \frac{1}{2}. \!
\]

With this definition of $\Delta$, and the assumption that $\mathcal{R}^\pi_{\gamma, \alpha} (T) = \Omega \left( T^{1 - \frac{\gamma (1 + \alpha)}{2 \gamma}} \right)$, for large enough $T$, $\tilde{C}_0 \Delta \leq \frac{1}{4}$.

Define the following function:

\[
\tilde{\phi}_0 (x) := \begin{cases} 
1/2 - \tilde{C}_0 x^{\gamma} & \text{if } 0 \leq x \leq \Delta^{\frac{1}{\gamma}} \\
1/2 - \tilde{C}_0 \left( 2\Delta^{\frac{1}{\gamma}} - x \right)^{\gamma} & \text{if } \Delta^{\frac{1}{\gamma}} \leq x \leq 2\Delta^{\frac{1}{\gamma}} \\
1/2 & \text{o.w.}
\end{cases}
\]

Let $\phi_0 (x) := \sum_{i=0}^{\nu - 1} \tilde{\phi}_0 (x - 2i\Delta^{\frac{1}{\gamma}})$, where $\nu = \lfloor \Delta^{\frac{\beta}{2}} \rfloor$; Clearly $\phi_0 \in \mathcal{H}(\gamma, L)$. For $1 \leq j \leq M \nu$, define the intervals $I_j := [4(j - 1)\Delta^{\frac{1}{\gamma}}, 4j\Delta^{\frac{1}{\gamma}}]$. Next, we construct some payoff functions $\phi_m, 1 \leq m \leq M$, that are in $\mathcal{H}(\beta, L)$ and are equal to $\phi_0$ everywhere except for some sections in $\tilde{I}_m := \bigcup_{1+(m-1)\nu \leq i \leq m \nu} I_i$. For $1 \leq i \leq M \nu$, define the following points

\[
x_i^- := (4i - 3)\Delta^{\frac{1}{\beta}}, \quad x_i := (4i - 2)\Delta^{\frac{1}{\beta}}, \quad x_i^+ := (4i - 1)\Delta^{\frac{1}{\beta}}, \quad \tilde{x}_i^- \leq x_i^- \quad \text{and} \quad \tilde{x}_i^+ \geq x_i^+ \quad \text{such that}
\]

\[
1/2 - \tilde{C}_0 (x_i^- - \tilde{x}_i^-)^{\beta} = \phi_0 (\tilde{x}_i^-), \quad 1/2 - \tilde{C}_0 (\tilde{x}_i^+ - x_i^+)^{\beta} = \phi_0 (\tilde{x}_i^+).
\]
Now, we are ready to define the functions $\phi_m, 1 \leq m \leq M$:

$$
\phi_m(x) :=
\begin{cases}
1/2 - \tilde{C}_0(x^-_i - x) & \text{if } \tilde{x}_i^- \leq x \leq x_i^- \text{ for some } 1 + (m - 1)\nu \leq i \leq m\nu \\
1/2 + \tilde{C}_0(x - x_i^-) & \text{if } x_i^- \leq x \leq x_i \text{ for some } 1 + (m - 1)\nu \leq i \leq m\nu \\
1/2 + \tilde{C}_0(x_i^+ - x) & \text{if } x_i \leq x \leq x_i^+ \text{ for some } 1 + (m - 1)\nu \leq i \leq m\nu \\
1/2 - \tilde{C}_0(x - x_i^+) & \text{if } x_i^+ \leq x \leq \tilde{x}_i^+ \text{ for some } 1 + (m - 1)\nu \leq i \leq m\nu \\
\phi_0(x) & \text{otherwise}
\end{cases}
$$

**Step 4 (Verifying the margin condition).** We now check that the margin condition is satisfied with parameter $\alpha$ when $f_1 = \phi_m$ and $f_2 = 1/2$ for all $0 \leq m \leq M$. For $m = 0$ and $\delta \leq \tilde{C}_0\Delta$, one has

$$
P_X \{0 < |\phi_0(X) - 1/2| \leq \delta\} \leq 2\nu\bar{\rho}\tilde{C}_0^{3/4}\delta^{3/4} \leq 2\Delta^{3/4} - \frac{1}{7} \bar{\rho}\tilde{C}_0^{1/2} \delta^{1/2} \leq 2\bar{\rho}\tilde{C}_0^{3/4} \delta^{3/4} \leq C_0\delta^{3/4},
$$

where we have used $\alpha \leq \frac{1}{7}$ in (a). For $m = 0$ and $\delta > \tilde{C}_0\Delta$, one has

$$
P_X \{0 < |\phi_0(X) - 1/2| \leq \delta\} \leq 2\nu\bar{\rho}\Delta^{3/4} \leq 2\bar{\rho}\tilde{C}_0^{3/4} \delta^{3/4} \leq C_0\delta^{3/4},
$$

For $1 \leq m \leq M$ and $\delta \leq \tilde{C}_0\Delta$, one has

$$
P_X \{0 < |\phi_m(X) - 1/2| \leq \delta\} \leq \sum_{i=1+(m-1)\nu}^{m\nu} \left[ P_X \{0 < |\phi_m(X) - 1/2| \leq \delta, \tilde{x}_i^- \leq X \leq x_i^-\} \right. \\
+ P_X \{0 < |\phi_m(X) - 1/2| \leq \delta, x_i^- \leq X \leq x_i\} \\
+ P_X \{0 < |\phi_m(X) - 1/2| \leq \delta, x_i \leq X \leq x_i^+\} \\
+ P_X \{0 < |\phi_m(X) - 1/2| \leq \delta, x_i^+ \leq X \leq \tilde{x}_i^+\} \right] \\
+ P_X \{0 < |\phi_0(X) - 1/2| \leq \delta\} \\
\leq 4\nu\bar{\rho}\tilde{C}_0^{3/4}\delta^{3/4} + 2\nu\bar{\rho}\tilde{C}_0^{3/4}\delta^{3/4} \leq \bar{\rho}\left(4\tilde{C}_0^{3/4} + 2\tilde{C}_0^{3/4}\right)\Delta^{3/4} - \frac{1}{7} \delta^{1/2} \\
\leq \bar{\rho}\left(4\tilde{C}_0^{3/4} + 2\tilde{C}_0^{3/4}\right) \delta^{3/4} \leq C_0\delta^{3/4},
$$

where we have used $\alpha \leq \frac{1}{7}$ in (a). The case $1 \leq m \leq M$ and $\delta > \tilde{C}_0\Delta$ can be analyzed similar to the case $m = 0$ and $\delta > \tilde{C}_0\Delta$. 

39
Step 5 (Event of interest). Define \( Q_{i,T} := \sum_{t=1}^{T} \mathbb{1} \{ x_i^{-} \leq X_t \leq x_i^{+} \} \) to be the number of times contexts fall into the interval \([x_i^{-}, x_i^{+}]\) during the entire time horizon. Define the event

\[
\mathcal{A} := \left\{ \exists 1 \leq i \leq M \nu : Q_{i,T} < \frac{\rho T \Delta^\frac{1}{3}}{3} \right\}
\]

to be the event on which the number of contexts that have fallen into the interval \([x_i^{-}, x_i^{+}]\) is less than \(\frac{\rho T \Delta^\frac{1}{3}}{3}\) for at least one \(i\). Note that

\[
\mathbb{P} \{ \mathcal{A} \} \leq \sum_{i=1}^{M \nu} \mathbb{P} \left\{ Q_{i,T} < \frac{\rho T \Delta^\frac{1}{3}}{3} \right\}.
\]

Define \( Z_t := \mathbb{1} \{ x_i^{-} \leq X_t \leq x_i^{+} \} \). In order to bound each of the summands on the right hand side of the above inequality, one can apply Bernstein’s inequality to \( Q_i \) by noting that \( \mathbb{E} Z_t \geq \frac{\rho T \Delta^\frac{1}{3}}{3}, |Z_t| \leq 1 \), and \( \text{Var} Z_t \leq \mathbb{E} Z_t^2 \leq \frac{\rho T \Delta^\frac{1}{3}}{3} \) to get

\[
\mathbb{P} \{ \mathcal{A} \} \leq M \nu \exp \left( -\frac{\rho T \Delta^\frac{1}{3}}{3}/5 \right)
\]

\[
\leq \left[ \mathcal{S}^\pi_{\gamma,\alpha}(T) \right]^{\alpha / (\gamma + 2 \beta / \gamma)} \exp \left( -\frac{\rho T \left[ \mathcal{S}^\pi_{\gamma,\alpha}(T) \right]^{-\gamma / (\gamma + 2 \beta / \gamma)}}{5} \right)
\]

\[
\leq T^{\frac{1}{2\gamma}} \exp \left( -\frac{\rho T \left[ \mathcal{S}^\pi_{\gamma,\alpha}(T) \right]^{-\gamma / (\gamma + 2 \beta / \gamma)}}{5} \right)
\]

\[
\leq T^{\frac{1}{2\gamma}} \exp \left( -\frac{4\gamma + 1}{2\gamma} \log T \right)
\]

\[
= T^{-2}
\]

for large enough \( T \), where (a) follows from \( \frac{(1-\alpha)^\gamma}{\gamma + 2 \beta / \gamma} \leq \frac{1}{2\gamma} \) for \( 0 \leq \alpha \leq \frac{1}{2} \) and \( \beta \leq \gamma \leq \frac{1}{2} \), and (b) holds by the assumption that \( \mathcal{R}^\pi_{\gamma,\alpha}(T) = \Omega \left( T^{1-\frac{\gamma(\gamma + 2 \beta / \gamma)}{2\gamma + 1}} \right) \) and \( 0 < \beta < \frac{\gamma}{2 - \alpha \gamma} \), which implies

\[
T \left[ T^{\frac{1}{\gamma + 2 \beta / \gamma}} \mathcal{R}^\pi_{\gamma,\alpha}(T) \right]^{\alpha / (\alpha + 1)} = \Omega \left( T^{1-\frac{\gamma(2-\alpha)\gamma + 1)}{(\gamma + 2 \beta / \gamma - 3)(2 \gamma + 1)} \right) \geq \varepsilon \log T
\]

for some constant \( \varepsilon > 0 \) and large enough \( T \). Denote by \( \tilde{\mathcal{S}}^\pi(T) := \mathbb{E}^\pi \left[ \sum_{t=1}^{T} \mathbb{1} \{ f_{x_i}(X_t) \neq f_{x_i}(X_t) \} \middle| \bar{\mathcal{A}} \right] \) the inferior sampling rate of policy \( \pi \) when the event \( \mathcal{A} \) fails. Note that

\[
(1 - \mathbb{P} \{ \mathcal{A} \}) \mathcal{S}^\pi(T) \leq \tilde{\mathcal{S}}^\pi(T) \leq \mathcal{S}^\pi(T) + T\mathbb{P} \{ \mathcal{A} \}.
\]

The last two displays together imply that

\[
|\mathcal{S}^\pi(T) - \tilde{\mathcal{S}}^\pi(T)| \leq cT^{-1}.
\]

(A.13)

This inequality implies that \( \mathcal{S}^\pi(T) \leq \tilde{\mathcal{S}}^\pi(T) + cT^{-1} =: \mathcal{S}^\pi_{\gamma,\alpha}(T) \) for any problem instance in \( P(\gamma, \alpha, 1) \).
For the rest of the proof, all the probabilities and expectations will be computed conditional on the failure of the event $A$.

**Step 6 (Choosing a single problem with smoothness $\beta$).** Let $N_{m,T} := \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = 1, X_t \in \tilde{I}_m \}$ be the number of times policy $\pi$ pulls arm 1 when covariates fall into the interval $\tilde{I}_m$. Note that by definition $E_{\pi,\varphi_0}^T \left[ \sum_{m=1}^{M} N_{m,T} \mid \bar{A} \right] \leq \bar{S}_{\pi,\alpha}(T)$, which implies that there exists some $1 \leq m^* \leq M$ such that

$$E_{\pi,\varphi_0}^T \left[ N_{m^*,T} \mid \bar{A} \right] \leq \frac{\bar{S}_{\pi,\alpha}(T)}{M} \leq \frac{S_{\pi,\alpha}(T)}{M} + cT^{-1},$$

where we have used (A.13) for the last inequality.

**Step 7 (Likelihood of distinguishing different smoothnesses).** In this step, we will show that policy $\pi$ cannot distinguish between $\phi_0$ and $\phi_{m^*}$ with a considerable probability. For any set of samples $\{(\pi_t, X_t, Y_{\pi_t,t})\}_{t=1}^{T}$, define the log-likelihood ratio $L_{m,T} = L_{m,T} \left( \{(\pi_t, X_t, Y_{\pi_t,t})\}_{t=1}^{T} \right)$ as

$$L_{m,T} := \sum_{t=1}^{T} \log \left( \frac{p_{\pi,\varphi_0} \{ Y_{\pi_t,t} \mid \pi_t, X_t \}}{p_{\pi,\varphi_m} \{ Y_{\pi_t,t} \mid \pi_t, X_t \}} \right)$$

$$\leq \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = 1, X_t \in \tilde{I}_m \} \cdot \left[ Y_{\pi_t,t} \log \left( \frac{\phi_0(X_t)}{\phi_m(X_t)} \right) + (1 - Y_{\pi_t,t}) \log \left( \frac{1 - \phi_0(X_t)}{1 - \phi_m(X_t)} \right) \right]$$

$$\leq \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = 1, X_t \in \tilde{I}_m \} \cdot \left[ Y_{\pi_t,t} \frac{(\phi_0(X_t) - \phi_m(X_t))}{\phi_m(X_t)} + (1 - Y_{\pi_t,t}) \frac{(\phi_m(X_t) - \phi_0(X_t))}{1 - \phi_m(X_t)} \right]$$

$$= \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = 1, X_t \in \tilde{I}_m \} \cdot \frac{(Y_{\pi_t,t} - \phi_m(X_t))(\phi_0(X_t) - \phi_m(X_t))}{\phi_m(X_t)(1 - \phi_m(X_t))},$$
where the last inequality follows from $\log(1 + x) \leq x$ for all $x > 0$. By taking expectations of the above inequality conditional on the event $\mathcal{A}$ for $m = m^*$, one obtains

$$
\mathbb{E}_{\pi, \phi_0} \left[ L_{m^*, T} \mid \mathcal{A} \right] \leq \mathbb{E}_{\pi, \phi_0} \left[ \sum_{t=1}^{T} \mathbbm{1} \{ \pi_t = 1, X_t \in \bar{I}_{m^*} \} \cdot \frac{(Y_{\pi_t, t} - \phi_{m^*}(X_t))(\phi_0(X_t) - \phi_{m^*}(X_t))}{\phi_{m^*}(X_t)(1 - \phi_{m^*}(X_t))} \mid \mathcal{A} \right] 
$$

$$
\leq \mathbb{E}_{\pi, \phi_0} \left[ \sum_{t=1}^{T} \mathbbm{1} \{ \pi_t = 1, X_t \in \bar{I}_{m^*} \} \cdot \frac{(Y_{\pi_t, t} - \phi_{m^*}(X_t))(\phi_0(X_t) - \phi_{m^*}(X_t))}{\phi_{m^*}(X_t)(1 - \phi_{m^*}(X_t))} \mid X_t \right] 
$$

$$
= \mathbb{E}_{\pi, \phi_0} \left[ \sum_{t=1}^{T} \mathbbm{1} \{ \pi_t = 1, X_t \in \bar{I}_{m^*} \} \cdot \frac{(\phi_0(X_t) - \phi_{m^*}(X_t))^2}{\phi_{m^*}(X_t)(1 - \phi_{m^*}(X_t))} \right] 
$$

$$
\leq \frac{64 \tilde{C}_0^2 \Delta^2}{3} \mathbb{E}_{\pi, \phi_0} \left[ N_{m^*, T} \mid \tilde{\mathcal{A}} \right] \quad (a) \leq \frac{64 \tilde{C}_0^2 \Delta^2 S^{\pi}_{\gamma, \alpha}(T)}{3M} + cT^{-1} \quad (c)
$$

for large enough $T$, where (a) follows from $\tilde{C}_0 \Delta \leq \frac{1}{4}$, (b) follows from the definition of $m^*$, and (c) holds by the definition of $\Delta$.

**Step 8 (Lower bounding the number of mistakes).** Let

$$
\tilde{N}_{m, T} := \sum_{t=1}^{T} \sum_{i=1+(m-1)\nu}^{m\nu} \mathbbm{1} \{ \pi_t = 1, X_t \in [x_i^-, x_i^+] \} 
$$

be the number of times policy $\pi$ pulls arm 1 when covariates fall into the interval $[x_i^-, x_i^+]$. We will use the following two lemmas in order to to show that with a strictly positive probability one has $\tilde{N}_{m^*, T} < \frac{\rho T V \Delta^{\frac{3}{2}}}{2}$ conditional on the event $\mathcal{A}$ under problem $m^*$. This will imply that policy $\pi$ makes at least $\frac{\rho T V \Delta^{\frac{3}{2}}}{2}$ number of mistakes under the problem $m^*$.

**Lemma A.2.** Let $\rho_0, \rho_1$ be two probability distributions supported on some set $\mathcal{X}$, with $\rho_0$ absolutely continuous with respect to $\rho_1$. Then for any measurable function $\Psi : \mathcal{X} \to \{0, 1\}$, one has:

$$
\mathbb{P}_{\rho_0} \{ \Psi(X) = 1 \} + \mathbb{P}_{\rho_1} \{ \Psi(X) = 0 \} \geq \frac{1}{2} \exp(-\text{KL}(\rho_0, \rho_1)).
$$

**Proof.** Define $\mathcal{B}$ to be the event that $\Psi(X) = 1$. One has

$$
\mathbb{P}_{\rho_0} \{ \Psi(X) = 1 \} + \mathbb{P}_{\rho_1} \{ \Psi(X) = 0 \} = \mathbb{P}_{\rho_0} \{ \mathcal{B} \} + \mathbb{P}_{\rho_1} \{ \overline{\mathcal{B}} \} \geq \int \min \{ d\rho_0, d\rho_1 \} \geq \frac{1}{2} \exp(-\text{KL}(\rho_0, \rho_1)),
$$

where the last inequality follows from [Tsybakov 2008] Lemma 2.6.

**Lemma A.3.** For any event $\mathcal{E} \in \mathcal{F}_{T^{-}} = \sigma(\pi_1, X_1, Y_{\pi_1,1}, \ldots, \pi_T, X_T, Y_{\pi_T,T})$ and arbitrary event $\mathcal{B}$, we
have
\[
\mathbb{E}_{\pi, \phi_0} \left[ L_{m,T} \mid \mathcal{E}, \mathcal{B} \right] \geq \log \left( \frac{\mathbb{P}_{\pi, \phi_0} \left\{ \mathcal{E} \mid \mathcal{B} \right\}}{\mathbb{P}_{\pi, \phi_m} \left\{ \mathcal{E} \mid \mathcal{B} \right\}} \right).
\]

Proof. The proof is a simple extension of the proof of Lemma 19 in [Kaufmann et al., 2016]; hence, it is omitted.

Denote by \( \rho_0 \) and \( \rho_m \) the distribution of \( \tilde{N}_{m,T} \) under the problems 0 and \( m \) conditional on the event \( \mathcal{A} \). Also, define the test function \( \Psi(x) = 1 \left\{ x \geq \frac{2 \nu T \Delta^2}{b} \right\} \). Under this choice of \( \rho_0 \), \( \rho_m \), and \( \Psi \), one can apply Lemma \[A.2\] to obtain
\[
\mathbb{P}_{\pi, \phi_0} \left\{ \tilde{N}_{m^*,T} \geq \frac{\rho T \nu \Delta^2}{2} \bigg| \bar{A} \right\} + \mathbb{P}_{\pi, \phi_m} \left\{ \tilde{N}_{m^*,T} < \frac{\rho T \nu \Delta^2}{2} \bigg| \bar{A} \right\} \geq \frac{1}{2} \exp(-\text{KL}(\rho_0, \rho_m)).
\]

In order to lower bound the right hand side of this inequality, we note that
\[
\mathbb{E}_{\pi, \phi_0} \left[ L_{m^*,T} \bigg| \mathcal{A} \right] = \sum_{s=1}^{T} \mathbb{E}_{\pi, \phi_0} \left[ L_{m^*,T} \bigg| \mathcal{A}, \tilde{N}_{m^*,T} = s \right] \mathbb{P}_{\pi, \phi_0} \left\{ \tilde{N}_{m^*,T} = s \bigg| \mathcal{A} \right\} \geq \sum_{s=1}^{T} \log \left( \frac{\mathbb{P}_{\pi, \phi_0} \left\{ \tilde{N}_{m^*,T} = s \bigg| \mathcal{A} \right\}}{\mathbb{P}_{\pi, \phi_m} \left\{ \tilde{N}_{m^*,T} = s \bigg| \mathcal{A} \right\}} \right) \mathbb{P}_{\pi, \phi_0} \left\{ \tilde{N}_{m^*,T} = s \bigg| \mathcal{A} \right\} = \text{KL}(\rho_0, \rho_m),
\]
where the inequality follows from Lemma \[A.3\] The last two displays along with \[A.14\] yield
\[
\mathbb{P}_{\pi, \phi_0} \left\{ \tilde{N}_{m^*,T} \geq \frac{\rho T \nu \Delta^2}{2} \bigg| \bar{A} \right\} + \mathbb{P}_{\pi, \phi_m} \left\{ \tilde{N}_{m^*,T} < \frac{\rho T \nu \Delta^2}{2} \bigg| \bar{A} \right\} \geq \frac{1}{2} \exp(-1).
\]

Now, we need to show that \( \mathbb{P}_{\pi, \phi_0} \left\{ \tilde{N}_{m^*,T} \geq \frac{\rho T \nu \Delta^2}{2} \bigg| \bar{A} \right\} \) is small. To do so, we apply Markov’s inequality as follows:
\[
\mathbb{P}_{\pi, \phi_0} \left\{ \tilde{N}_{m^*,T} \geq \frac{\rho T \nu \Delta^2}{2} \bigg| \bar{A} \right\} \leq \frac{\mathbb{E}_{\pi, \phi_0} \left[ \tilde{N}_{m^*,T} \bigg| \mathcal{A} \right]}{\frac{\rho T \nu \Delta^2}{2}} \leq \frac{\mathbb{E}_{\pi, \phi_0} \left[ N_{m^*,T} \bigg| \mathcal{A} \right]}{\frac{\rho T \nu \Delta^2}{2}} \leq \frac{\text{SN}_{\alpha}(T)}{M} + cT^{-1} \overset{(a)}{=} \mathcal{O} \left( \frac{\text{SN}_{\alpha}(T)}{M} \right)
\]
\[
\overset{(b)}{=} \mathcal{O} \left( T^{-1} \left[ S_{\alpha, m}(T) \right] \right)
\]
\[
\overset{(c)}{=} \mathcal{O} \left( T^{-1} \left[ \frac{2+\alpha}{\alpha+1}(\beta \gamma - \beta) \left[ R_{\alpha, m}(T) \right] \right] \right)
\]
\[
\overset{(d)}{=} \mathcal{O} \left( 1/\log T \right),
\]
where (a) follows from the definition of \(m^*\) and (A.13), (b) holds by the definition of \(\Delta\), (c) is true due to Lemma B.2 and (d) holds by the assumption that \(R^\pi_{\gamma,\alpha}(T) = \Omega\left(T^{1 - \frac{\gamma(1 + \alpha)}{2\gamma + 1}}\right)\) and \(0 < \beta < \frac{\alpha\gamma^2}{1 - 4\gamma^2}\), which implies \(T^{-1}\left[T^{\frac{1}{1 + \pi}}_\nu \left[R^\pi_{\gamma,\alpha}(T)\right]^{\alpha + 1}_{\alpha + 1}\right]^{(2\beta\gamma + \gamma - \beta)} = O\left(T^{1 - \frac{(\alpha\beta\gamma + \gamma - \beta)/((\gamma - \beta)(2\gamma + 1))}{2\gamma + 1}}\right) \leq \tilde{c}\log^{-1} T\) for some constant \(\tilde{c} > 0\) and large enough \(T\). The last two displays yield that for large enough \(T\), one has

\[
\mathbb{P}_{\pi, \phi, m^*} \left\{ \frac{\tilde{N}_{m^*, T}}{2} \right\} \geq \frac{1}{4e}.
\]

Note that by definition, when the event \(\tilde{A}\) holds, at least \(\tilde{\rho}T\nu\Delta^{\frac{1}{2}}\) number of contexts fall into the interval \(\bigcup_{1 + (m^* - 1)\nu \leq m^* \nu; x_i^{-}, x_i^{+}}\), that is,

\[
S^\pi(T; \phi, m^*) \geq \frac{\tilde{\rho}T\nu\Delta^{\frac{1}{2}}}{2} \mathbb{P}_{\pi, \phi, m^*} \left\{ \left\{ \tilde{N}_{m^*, T} < \frac{\tilde{\rho}T\nu\Delta^{\frac{1}{2}}}{2} \right\} \right\} \geq \frac{\tilde{\rho}T\nu\Delta^{\frac{1}{2}}}{8e} \geq C'T \left[R^\pi_{\gamma,\alpha}(T)\right]^{-\frac{\alpha\beta\gamma + \gamma - \beta}{2\beta\gamma + \gamma - \beta}},
\]

for some constant \(C' > 0\).

**Step 9 (From inferior sampling rate to regret).** Finally using the above inequality along with (A.13), and Lemma A.1 one obtains

\[
\sup \left\{ R^\pi(P; T) : P \in \mathcal{P}(\beta, \alpha, 1) \right\} \geq R^\pi(\phi, m^*; T) \geq \tilde{C}T^{1 - \frac{1}{2} - \frac{\alpha\beta\gamma + \gamma - \beta}{2\beta\gamma + \gamma - \beta}}\left[R^\pi_{\gamma,\alpha}(T)\right]^{-\frac{\alpha\beta\gamma + \gamma - \beta}{2\beta\gamma + \gamma - \beta}},
\]

for some constant \(\tilde{C} > 0\). This concludes the proof.

**A.5 Proof of Proposition 4.7**

Let \(\tilde{r} := \left[2\log_2\left(\frac{\gamma}{2C_3}\right) + 2l + \left(\frac{d}{2} + 1\right)\log_2\log T\right]\) where the constant \(C_3\) was introduced in Proposition 4.5. We will prove the result by bounding the following probability

\[
\mathbb{P}\left\{ \exists r \in [\tilde{r}] : \sup_{k \in \mathcal{K}, x \in \mathcal{M}(B)} \left| f^r_k(x; j^r_1(B)) - f^r_k(x; j^r_2(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{r/2}} \right\} \leq \sum_{r \in [\tilde{r}]} \sum_{k \in \mathcal{K}} \sum_{x \in \mathcal{M}(B)} \mathbb{P}\left\{ \left| f^r_k(x; j^r_1(B)) - f^r_k(x; j^r_2(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{r/2}} \right\}.
\]

Note that by the triangle inequality,

\[
\left| f^r_k(x; j^r_1(B)) - f^r_k(x; j^r_2(B)) \right| \leq \left| f_k(x) - f^r_k(x; j^r_1(B)) \right| + \left| f_k(x) - f^r_k(x; j^r_2(B)) \right| + \left| f^r_k(x; j^r_1(B)) - f^r_k(x; j^r_2(B)) \right|.
\]

44
That is,
\[
\begin{align*}
\mathbb{P} \left\{ \left| \hat{f}_k^\ell(B,r) (x; j_1(B)) - \hat{f}_k(B,r) (x; j_1(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1+r/2}} \right\} & \\
\leq \mathbb{P} \left\{ \left| f_k(x) - \hat{f}_k^\ell(B,r) (x; j_1(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1+r/2}} \right\} & \\
+ \mathbb{P} \left\{ \left| f_k(x) - \hat{f}_k(B,r) (x; j_2(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1+r/2}} \right\}. \quad \text{(A.16)}
\end{align*}
\]

Note that since when \( r \leq \tilde{r} \) we have \( \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1+r/2}} \geq C_3 2^{-\beta j_2} \geq C_3 2^{-\beta j_2} \), we can apply Proposition 4.5 to bound the two terms on the right hand side of above inequality. Namely, we apply Proposition 4.5 with \( n = 2^r, \mu = \frac{\rho}{\mu^2 \cdot \tilde{\rho}}, \tilde{\mu} = \frac{\rho}{\mu^2 \cdot \tilde{\rho}}, \delta = \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1+r/2}}, \) and \( h = 2^{-j_2} \) for the first term and \( h = 2^{-j_2} \) for the second term to obtain
\[
\begin{align*}
\mathbb{P} \left\{ \left| f_k(x) - \hat{f}_k^\ell(B,r) (x; j_1(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1+r/2}} \right\} & \leq \tilde{C}_1 T^{-\gamma^2 \tilde{C}_2}, \\
\mathbb{P} \left\{ \left| f_k(x) - \hat{f}_k(B,r) (x; j_2(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1+r/2}} \right\} & \leq \tilde{C}_1 T^{-\gamma^2 \tilde{C}_2},
\end{align*}
\]
where the constants \( \tilde{C}_1, \tilde{C}_2 \) depend only on \( \tilde{\beta}, L, \rho, \tilde{\rho}, \) and \( d \). These two inequalities along with (A.15) and (A.16) imply
\[
\mathbb{P} \left\{ \exists r \in [\tilde{r}, \bar{r}] : \sup_{k \in \mathcal{K}, x \in \mathcal{M}(B)} \left| \hat{f}_k^\ell(B,r) (x; j_1(B)) - \hat{f}_k(B,r) (x; j_2(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{r/2}} \right\} \leq 2 \left| \mathcal{M}(B) \right| (\bar{r} - \tilde{r}) \tilde{C}_1 T^{-\gamma^2 \tilde{C}_2}
\]
\[
\leq C_7 (\log T)^{\frac{d}{2}} T^{-\gamma^2 C_8 + C_9},
\]
where the constants \( C_7, C_8, C_9 \) depend on only \( \tilde{\beta}, \tilde{\beta}, L, \rho, \tilde{\rho}, \) and \( d \). The results follows by applying union bound over \( B \in \mathcal{B}_l \). This concludes the proof. \( \square \)

### A.6 Proof of Proposition 4.8

First, we need to determine the open cube \( B \in \mathcal{B}_l \) for which, we want to prove the inequality. Fix arm \( k \in \mathcal{K} \). Recall that by Assumption 4.4 there exists at least one open cube \( \tilde{B} \in \mathcal{B}_l \) and a point \( \hat{x} \in \tilde{B} \) such
that
\[ \left| \mathbf{I}_{j_1}^\beta f_k(\hat{x}; \tilde{B}) - f_k(\hat{x}) \right| = \left| \mathbf{I}_{j_1}^{[\beta]} f_k(\hat{x}) - f_k(\hat{x}; \tilde{B}) \right| \geq b2^{-t \beta}. \quad (A.17) \]

Let \( \hat{x} = \arg \min_{x \in M(\theta)} \|x - \hat{x}\|_\infty \) (if there is more than one minimizer we choose the one with the minimum \( L_1 \)-norm). Note that \( \|\tilde{x} - \hat{x}\|_\infty \leq 2^{-t} \), which along with the assumption \( f_k \in \mathcal{H}(\beta, L) \) implies that
\[ \left| f_k(\hat{x}) - f_k(\hat{x}) \right| \leq L \|\tilde{x} - \hat{x}\|_\infty \leq L 2^{-t \beta} \leq \frac{L}{\log T} 2^{-t \beta}. \quad (A.18) \]

In addition, by Lemma 4.1 we have
\[ \left| \mathbf{I}_{j_1}^{[\beta]} f_k(\hat{x}; \tilde{B}) - \mathbf{I}_{j_1}^{[\beta]} f_k(\hat{x}; \tilde{B}) \right| \leq \kappa_0 2^t \|\tilde{x} - x\|_\infty \leq \kappa_0 2^t \leq \frac{\kappa_0}{\log T} 2^{-t \beta}, \quad (A.19) \]

where \( \kappa_0 \) was introduced in Lemma 4.1. Let \( \tilde{r} := [2 \log_2(\frac{4\gamma}{L \kappa_0}) + 2\beta + (\frac{d}{2} + 3) \log_2 \log T] \). We have
\[ \mathbb{P} \left\{ r_{\text{last}}(\tilde{B}) > \tilde{r} \right\} \leq \mathbb{P} \left\{ \left| j_k(\tilde{B}, r) (\tilde{x}; j_1(\tilde{B})) - j_k(\tilde{B}, r) (\tilde{x}; j_2(\tilde{B})) \right| < \frac{\gamma (\log T) \frac{d}{2} + \frac{1}{2}}{2^{\frac{d}{2}}} \right\}. \quad (A.20) \]

Note that by the triangle inequality,
\[ \left| j_k(\tilde{B}, r) (\tilde{x}; j_1(\tilde{B})) - j_k(\tilde{B}, r) (\tilde{x}; j_2(\tilde{B})) \right| \geq \left| f_k(\tilde{x}) - f_k(\tilde{B}, r) (\tilde{x}; j_2(\tilde{B})) \right| > \left| f_k(\tilde{x}) - f_k(\tilde{B}, r) (\tilde{x}; j_1(\tilde{B})) \right|. \quad (A.21) \]

Note that since we have \( \frac{\gamma (\log T) \frac{d}{2} + \frac{1}{2}}{2^{\frac{d}{2}}} \geq C_3 2^{-j_2^B} \), we can apply Proposition 4.5 to show that second term on the right hand side of above inequality is “small” with high probability. Namely, we apply Proposition 4.5 with \( n = 2^\hat{r} \), \( \mu = \frac{\rho}{\rho_2 - \rho} \), \( \bar{\mu} = \frac{\bar{\rho}}{\rho_2 - \rho} \), \( \delta = \frac{\gamma (\log T) \frac{d}{2} + \frac{1}{2}}{2^{\frac{d}{2}}} \), and \( h = 2^{-j_2^B} \) to obtain
\[ \mathbb{P} \left\{ \left| f_k(\tilde{x}) - f_k(\tilde{B}, r) (\tilde{x}; j_2(\tilde{B})) \right| \geq \frac{\gamma (\log T) \frac{d}{2} + \frac{1}{2}}{2^{\frac{d}{2}}} \right\} \leq C_1 T^{-\gamma^2 C_2}, \quad (A.22) \]

where the constants \( C_1, C_2 \) depend only on \( \beta, L, \rho, \bar{\rho}, d \). Now, we show that the first term on the right hand side of (A.20) cannot get “small” with high probability. One can write
\[ \left| f_k(\tilde{x}) - f_k(\tilde{B}, r) (\tilde{x}; j_1(\tilde{B})) \right| \geq \left| f_k(\tilde{x}) - f_k(\tilde{B}, r) (\tilde{x}; j_1(\tilde{B})) \right| - \left| \mathbf{I}_{j_1}^{[\beta]} f_k(\tilde{x}; \tilde{B}) - f_k(\tilde{B}, r) (\tilde{x}; j_1(\tilde{B})) \right|. \quad (A.23) \]
The first corresponds to bias and the second term corresponds to stochastic error. Note that by (A.17), (A.18), and (A.19), we have
\[ \left| f_k(\tilde{x}) - \Gamma_{j_1}^{[\beta]} f_k(\tilde{x}; \bar{B}) \right| \geq \left| f_k(\tilde{x}) - \Gamma_{j_1}^{[\beta]} f_k(\tilde{x}; \bar{B}) \right| - \left| f_k(\tilde{x}) - f_k(\hat{x}) \right| - \left| \Gamma_{j_1}^{[\beta]} f_k(\hat{x}; \bar{B}) - \Gamma_{j_1}^{[\beta]} f_k(\tilde{x}; \bar{B}) \right| \]
\[ \geq b 2^{-1/\beta} - \frac{L}{\log T} 2^{-1/\beta} - \frac{\kappa_0}{\log T} 2^{-1/\beta} \geq \frac{2 \gamma (\log T)^{\frac{d}{2}\gamma + \frac{1}{2}}}{2^{\gamma/2}} \]
for large enough $T \geq T_0(L, b, \rho, \tilde{\rho}, d)$. In order to bound the second term on the right hand side of (A.22), we apply Proposition 4.6 with $n = 2^\beta$, $\delta = \frac{(\log T)^{\frac{d}{2}\gamma + \frac{1}{2}}}{2^{1+\gamma/2}}$, and $h = 2^{-j_1\beta}$ to obtain
\[ \mathbb{P} \left\{ \left| \Gamma_{j_1}^{[\beta]} f_k(\tilde{x}; \bar{B}) - f_k(\hat{x}, \bar{B}) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2}\gamma + \frac{1}{2}}}{2^{1+\gamma/2}} \right\} \leq \tilde{C}_4 T^{-\gamma^2 \tilde{C}_5}, \]
where the constants $\tilde{C}_4, \tilde{C}_5$ depend only on $\tilde{\beta}, L, \rho, \tilde{\rho}$, and $d$. Putting together (A.20), (A.21), (A.23), and (A.24), one obtains
\[ \mathbb{P} \left\{ r_{\text{last}}^{(B)} > \hat{r} \right\} \leq \mathbb{P} \left\{ \left| \Gamma_{j_1}^{[\beta]} f_k(\tilde{x}; \bar{B}) - f_k(\hat{x}, \bar{B}) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2}\gamma + \frac{1}{2}}}{2^{1+\gamma/2}} \right\} + \mathbb{P} \left\{ \left| \Gamma_{j_1}^{[\beta]} f_k(\tilde{x}; \bar{B}) - f_k(\hat{x}, \bar{B}) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2}\gamma + \frac{1}{2}}}{2^{1+\gamma/2}} \right\} \leq C_{10} T^{-\gamma^2 C_{11}}, \]
where the last inequality follows from (A.22) and (A.25), and where the constants $C_{10}, C_{11}$ depend only on $\tilde{\beta}, L, \rho, \tilde{\rho}$, and $d$. This concludes the proof.

**A.7 Proof of Corollary 4.9**

Note that for large enough $T$, one has
\[ \mathbb{P} \left\{ \beta_{SACB} \in [\beta - \frac{3(\beta^2 + d) 2 \log_2 \log T}{(\beta + d - 1) \log_2 T}, \beta] \right\} \leq \mathbb{P} \left\{ 2l \beta + \left( \frac{d}{\beta} + 1 \right) \log_2 \log T \leq r_{\text{last}}^{(B)} \leq 2l \beta + \left( \frac{d}{\beta} + 4 \right) \log_2 \log T \right\} \leq 1 - C_7 2^d (\log T)^{\frac{d}{2}} T^{-\gamma^2 C_8 + C_9} - C_10 T^{-\gamma^2 C_{11}}, \]
where the last inequality follows from Propositions 4.7 and 4.8 and the constants $C_7, C_8, C_9 > 0$ were introduced in the former proposition, and the constants $C_{10}, C_{11} > 0$ were introduced in the latter. This concludes the proof.

A.8 Proof of Theorem 4.10

We only discuss the case $\beta \leq 1$. The case $\beta > 1$ can be analyzed similarly. First, we show that with high probability, $T_{\text{SACB}} \leq \frac{4}{\rho} (\log T)^{\frac{2d}{\beta} + 4} T^\frac{(\beta + d - 1) \log_2 T}{(2\beta + d)^2} =: \bar{T}_{\text{SACB}}$. Note that the policy terminates when all the open cubes $B \in B_l$ have reached round $\bar{r} = \lceil 2l\bar{\beta} + (\frac{2d}{\beta} + 4) \log_2 T \rceil$. That is, $T_{\text{SACB}}$ is less than the time step at which $2 \sum_{r=1}^{\bar{r}} 2^r$ number of contexts have fallen into each $B \in B_l$. Note that

$$\sum_{r=1}^{\bar{r}} 2^r \leq 2^{\bar{r}+1} \leq 2 (\log T)^{\frac{2d}{\beta} + 4} T^\frac{2\bar{r}(d + \beta d - 1)}{(2\beta + d)^2}.$$

Let $\bar{N}(B) := \sum_{t=1}^{T_{\text{SACB}}} Z_t$ be the number of contexts that have fallen into $B$ by $t = \bar{T}_{\text{SACB}}$, where $Z_t$’s are i.i.d Bernoulli random variables with $E[Z_t] \geq \rho T^\frac{d}{(2\beta + d)^2}$ and $\text{Var}(Z_t) \leq \rho T^\frac{d}{(2\beta + d)^2}$. Now, we can apply the Bernstein’s inequality in Lemma B.2 to $\bar{N}(B)$ with $a = 2 (\log T)^{\frac{2d}{\beta} + 4} T^\frac{2\bar{r}(d + \beta d - 1)}{(2\beta + d)^2}$, which yields

$$\mathbb{P}\left\{ \bar{N}(B) < 2 (\log T)^{\frac{2d}{\beta} + 4} T^\frac{2\bar{r}(d + \beta d - 1)}{(2\beta + d)^2} \right\} \leq \exp\left(\frac{-a^2}{2\bar{T}_{\text{SACB}} \text{Var}(Z_t) + a}\right) \leq \exp\left(\frac{-\rho}{4\rho + 2\rho} (\log T)^{\frac{2d}{\beta} + 4} T^\frac{2\bar{r}(d + \beta d - 1)}{(2\beta + d)^2}\right),$$

and, by the union bound

$$\mathbb{P}\{ T_{\text{SACB}} > \bar{T}_{\text{SACB}} \} \leq \sum_{B \in B_l} \mathbb{P}\left\{ \bar{N}(B) < 2 (\log T)^{\frac{2d}{\beta} + 4} T^\frac{2\bar{r}(d + \beta d - 1)}{(2\beta + d)^2} \right\} \leq 2T^\frac{d}{(2\beta + d)^2} \exp\left(\frac{-\rho}{4\rho + 2\rho} (\log T)^{\frac{2d}{\beta} + 4} T^\frac{2\bar{r}(d + \beta d - 1)}{(2\beta + d)^2}\right).$$
As a result, the regret incurred up to $t = \lceil T_{\text{SACB}} \rceil$ is bounded by

\[
\mathbb{E}^\pi \left[ \sum_{t=1}^{\lceil T_{\text{SACB}} \rceil} f_{\pi_t}(X_t) - f_{\pi_t}(X_t) \right] \leq T \cdot \mathbb{P} \{ T_{\text{SACB}} > T_{\text{SACB}} \} + T_{\text{SACB}} \leq 2T^{1+\frac{d(\beta+d-1)}{(2\beta+d)^2}} \exp \left( -\frac{\rho}{4\rho + 2\rho} (\log T)^{\frac{2d}{2} + 4} \frac{T}{(2\beta+d)^2} \right) + \frac{4}{\rho} (\log T)^{\frac{2d}{2} + 4} \frac{T}{(2\beta+d)^2}.
\]

where the last equality follows from $\frac{d(\beta+d-1)}{(2\beta+d)^2} \leq 1 - \frac{d(\alpha+1)}{2\beta+d}$ for any $\beta \leq \beta \leq \beta$ and $\alpha \leq \frac{1}{T^2}$. Define $\hat{\beta}_T := \beta - \frac{3(2\beta+d)^2 \log_T \log_T}{(\beta+d-1) \log_2 T}$. Finally, the regret from $t = \lceil T_{\text{SACB}} \rceil + 1$ to $T = T$ is bounded by

\[
\mathbb{E}^\pi \left[ \sum_{t=\lceil T_{\text{SACB}} \rceil + 1}^{T} f_{\pi_t}(X_t) - f_{\pi_t}(X_t) \right] \leq T \cdot \mathbb{P} \{ \hat{\beta}_{\text{SACB}} \notin [\hat{\beta}_T, \beta] \} + C T^{1 - \frac{\beta(\alpha+1)}{2\beta+d}} \log T \leq C_{12} (\log T)^{\frac{d}{2}} T^{\gamma^2 C_{13} + C_{14} + 1 + \frac{d(\beta+d-1)}{(2\beta+d)^2}} + C T^{1 - \frac{\beta(\alpha+1)}{2\beta+d}} (\log T)^{\frac{3d(\alpha+1)(2\beta+d)^2}{(2\beta+d)^2}}.
\]

where the constants $C_{12}, C_{13},$ and $C_{14}$ were introduced in Theorem 4.9. Putting together (A.26) and (A.27) concludes the proof.

\[\Box\]

### A.9 Proof of Proposition 5.3

For simplicity of notation, let $\tilde{\beta} = \min_k \beta_{f_k}(B)$ and $k = \arg \min_m \beta_{f_m}(B)$. The lower bound can be established by following the same exact lines of proof for Proposition 4.7, the only difference is that one needs to use $\bar{r} := \lceil 2 \log_T (\frac{2\beta}{\beta_\alpha}) + (l^{(B)} + 3)\beta + \frac{2d}{2} + 1 \rceil \log_T T$, and to replace $l$ with $l^{(B)}$, and $\beta$ with $\hat{\beta}$.

In order to prove the upper bound Recall that by Assumption 6 there exists at least a point $\hat{x} \in B$ and an arm $k \in K$ such that

\[
\left| T_{j_1}^{0}(B) f_k(\hat{x}; B) - f_k(\hat{x}) \right| \geq \frac{2f^{(B)}_1}{\log T}. \tag{A.28}
\]

Let $\bar{x} = \arg \min_{x \in \mathcal{M}(\hat{B})} \| x - \hat{x} \|_\infty$ (if there is more than one minimizer we choose the one with the minimum $L_1$-norm). Note that $\| \bar{x} - \hat{x} \|_\infty \leq 2^{-l^{(B)}}$, which along with the assumption $f_k \in \mathcal{H}(\hat{\beta}, L)$ implies
that
\[ |f_k(\tilde{x}) - f_k(\hat{x})| \leq L\|\tilde{x} - \hat{x}\|_\infty \leq L2^{-\tilde\gamma(B)\tilde{\beta}} \leq \frac{L2^{-j_1(B)\tilde{\beta}}}{\log^2 T}. \] (A.29)

In addition, by Lemma 4.1, we have
\[ |\Gamma^0 f_k(\tilde{x}; B) - \Gamma^0 f_k(\hat{x}; B)| \leq \kappa_0 2^{l(B)} \|\tilde{x} - x\|_\infty \leq \kappa_0 2^{l(B) - \tilde{l}(B)} \leq \frac{\kappa_0 2^{-j_1(B)\tilde{\beta}}}{\log^2 T}, \] (A.30)

where \( \kappa_0 \) was introduced in Lemma 4.1. Let \( \tilde{r} := [2 \log_2(\frac{4\gamma}{L\kappa_0}) + 2(l(B) + 3)\tilde{\beta} + (\frac{2d}{3} + 5)\log_2 \log T] \). We have
\[ \mathbb{P}\left\{ \text{r}_{\text{last}} > \tilde{r} \right\} \leq \mathbb{P}\left\{ \left| \hat{f}^B_k(\tilde{x}; j_1(B)) - \hat{f}^B_k(\tilde{x}; j_2(B)) \right| < \frac{\gamma (\log T) 2^{\tilde{\beta}_2/2}}{2^{r/2}} \right\}. \] (A.31)

Note that by the triangle inequality,
\[ \left| \hat{f}^B_k(\tilde{x}; j_1(B)) - \hat{f}^B_k(\tilde{x}; j_2(B)) \right| \geq \left| f_k(\tilde{x}) - \hat{f}^B_k(\tilde{x}; j_1(B)) \right| - \left| f_k(\tilde{x}) - \hat{f}^B_k(\tilde{x}; j_2(B)) \right|. \] (A.32)

Note that since we have \( \frac{\gamma (\log T) 2^{\tilde{\beta}_2}}{2^{r/2}} \geq C_3 2^{-j_2^B} \), we can apply Proposition 4.5 to show that second term on the right hand side of above inequality is “small” with high probability. Namely, we apply Proposition 4.5 with \( n = 2^r \), \( \mu = \frac{\rho}{\rho_2 - d[\tilde{\beta}]} \), \( \bar{\mu} = \frac{\rho}{\rho_2 - d[\tilde{\beta}]} \), \( \beta = \frac{\gamma (\log T) 2^{\tilde{\beta}_2}}{2^{r/2}} \), and \( h = 2^{-j_2^B} \) to obtain
\[ \mathbb{P}\left\{ \left| f_k(\tilde{x}) - \hat{f}^B_k(\tilde{x}; j_1(B)) \right| \geq \frac{\gamma (\log T) 2^{\tilde{\beta}_2/2}}{2^{r/2}} \right\} \leq \tilde{C}_1 T^{-\gamma^2 \tilde{C}_2}, \] (A.33)

where the constants \( \tilde{C}_1, \tilde{C}_2 \) depend only on \( \tilde{\beta}, L, \rho, \tilde{\rho}, \) and \( d \). Now, we show that the first term on the right hand side of (A.31) cannot get “small” with high probability. One can write
\[ \left| f_k(\tilde{x}) - \hat{f}^B_k(\tilde{x}; j_1(B)) \right| \geq \left| f_k(\tilde{x}) - \Gamma^0_{j_1(B)} f_k(\tilde{x}; B) \right| - \left| \Gamma^0_{j_1(B)} f_k(\tilde{x}; B) - \hat{f}^B_k(\tilde{x}; j_1(B)) \right|. \] (A.34)

The first corresponds to bias and the second term corresponds to stochastic error. Note that by (A.28), (A.29), and (A.30), we have
\[ \left| f_k(\tilde{x}) - \Gamma^0_{j_1(B)} f_k(\tilde{x}; B) \right| \geq \left| f_k(\tilde{x}) - \Gamma^0_{j_1(B)} f_k(\tilde{x}; B) \right| - \left| f_k(\tilde{x}) - f_k(\tilde{x}) \right| - \left| \Gamma^0_{j_1(B)} f_k(\tilde{x}; B) - \Gamma^0_{j_1(B)} f_k(\tilde{x}; B) \right| \geq 2^{-j_1(B)\tilde{\beta}} \frac{L2^{-j_1(B)\tilde{\beta}}}{\log^2 T} \geq 2^{-j_1(B)\tilde{\beta}} \frac{\kappa_0 2^{-j_1(B)\tilde{\beta}}}{\log^2 T} \geq \frac{L \wedge \frac{\kappa_0 2^{-j_1(B)\tilde{\beta}}}{\log^2 T}}{2^{1/2}} \geq \frac{\kappa_0 2^{-j_1(B)\tilde{\beta}}}{\log^2 T} \frac{\gamma (\log T) 2^{\tilde{\beta}_2/2}}{2^{r/2}} \] (A.35)

for large enough \( T \geq T_0(L, \rho, \tilde{\rho}, d) \). In order to bound the second term on the right hand side of (A.33),
we apply Proposition \[4.6\] with \(n = 2^{j_0}\), \(\delta = \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1 + r/2}}\), and \(h = 2^{-j_0}\) to obtain
\[
P \left\{ \left| \Gamma_{j_1(B)} f_k(\tilde{x}; B) - \tilde{f}_{j_1}(\tilde{x}; j_1(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1 + r/2}} \right\} \leq \tilde{C}_4 T^{-\gamma^2 \tilde{C}_5}, \tag{A.36} \]
where the constants \(\tilde{C}_4, \tilde{C}_5\) depend only on \(\tilde{\beta}, L, \rho, \bar{\rho}\), and \(d\). Putting together (A.31), (A.32), (A.34), and (A.35), one obtains
\[
P \left\{ r_{\text{last}}(B) > \hat{r} \right\} \leq \frac{3}{2} \gamma^2 \frac{C_{11}}{2^{1/2}} (\log T)^{11/2} \leq \left\{ \left| \Gamma_{j_1(B)} f_k(\tilde{x}; B) - \tilde{f}_{j_1}(\tilde{x}; j_1(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1 + r/2}} \right\}
+ \frac{3}{2} \gamma^2 \frac{C_{11}}{2^{1/2}} (\log T)^{11/2} \leq \left\{ \left| \Gamma_{j_1(B)} f_k(\tilde{x}; B) - \tilde{f}_{j_1}(\tilde{x}; j_1(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1 + r/2}} \right\}
\leq C_{10} T^{-\gamma^2 C_{11}},
\]
where the last inequality follows from (A.33) and (A.36), and where the constants \(C_{10}, C_{11}\) depend only on \(\tilde{\beta}, L, \rho, \bar{\rho}, \) and \(d\). This concludes the proof.

\[\square\]

A.10 Proof of Proposition 5.4

\textbf{Part a.} Assume \(\sup_{x \in B} f_1(x) - f_2(x) \geq \frac{3}{2} \gamma^2 \frac{C_{0,3}}{2^{1/2}} (\log T)^{11/2} \) \(\beta_{k,f_k(B)}\), and let \(\tilde{x} \in B \cup \partial B\) be the point in \(B\) or its boundary, at which the supremum is attained. On the event \(S(B)\), one has
\[
P \left\{ \Delta(B) < \frac{\gamma (\log T)^{d/2}}{2^{r_{\text{last}}(B)/2}} \right\} \leq \frac{3}{2} \gamma^2 \frac{C_{0,3}}{2^{1/2}} (\log T)^{11/2} \leq \left\{ \left| \Gamma_{j_1(B)} f_k(\tilde{x}; B) - \tilde{f}_{j_1}(\tilde{x}; j_1(B)) \right| \geq \frac{\gamma (\log T)^{\frac{d}{2} + \frac{1}{2}}}{2^{1 + r/2}} \right\}
\leq \sum_{k \in K} \mathbb{P} \left\{ \left| \Gamma_{j_1(B)} f_k(\tilde{x}) - \omega^{(B)}(\tilde{x}; D_{k, j_1}^{(B), \text{last} + 1}(\tilde{x})) \right| \geq \gamma \frac{C_{0,3}}{2^{1/2}} (\log T)^{11/2} \right\}
= \sum_{k \in K} \mathbb{P} \left\{ \left| f_k(\tilde{x}) - \hat{L}(\tilde{x}; D_{k, j_1}^{(B), \text{last} + 1}(\tilde{x}}, 2^{-l(B)} \right| \geq \gamma \frac{C_{0,3}}{2^{1/2}} (\log T)^{11/2} \right\}.
\]
Note that since we have
\[
\frac{3}{2} \gamma^2 \frac{C_{0,3}}{2^{1/2}} (\log T)^{11/2} \geq C_3 \frac{2^{-L_{k,f_k(B)}}}{2^{l(B)} \min_{\beta_{k,f_k(B)}}},
\]
we can apply Proposition \[4.5\] to show that last term on the right hand side of the above display is “small.” Namely, we apply Proposition \[4.5\]
with \( n = 2^{r_{\text{last}}} \), \( \mu = \frac{\bar{p}}{\rho_2 - d_{\text{left}}} \), \( \hat{\mu} = \frac{\bar{p}}{\rho_2 - d_{\text{left}}} \), \( \delta = \frac{\gamma_2 2^{r_{\text{last}} + 3\beta}}{2^{r_{\text{last}}} \min_k \beta_k(B)} \), and \( h = 2^{-\beta} \) to obtain

\[
\sum_{k \in K} P \left\{ f_k(\tilde{x}) - \eta \Delta \left( \tilde{x} ; D^{(B,r_{\text{last}}+1)}_k , 2^{-l} \right) \geq \gamma \frac{2^{r_{\text{last}} + 3\beta}}{2^{r_{\text{last}}} \min_k \beta_k(B)} \right\} \leq C_7 T^{-\gamma^2 C_{18}}, \quad (A.37)
\]

where \( C_7 \) and \( C_{18} \) that depend on only \( \beta, L, \rho, \bar{p}, \) and \( d \).

**Part b.** The proof is similar to the previous part and hence omitted.

**Part c.** Assume \( \sup_{x \in B} |f_1(x) - f_2(x)| \leq \frac{\gamma_2 2^{r_{\text{last}} + 3\beta}}{3 \times 2^{r_{\text{last}}} \min_k \beta_k(B)} \), and let \( \tilde{x} \in B \cup \partial B \) be the point in \( B \) or its boundary, at which the supremum is attained. On the event \( \mathcal{S}(B) \), one has

\[
P \left\{ \left| \Delta(B) \right| \geq \gamma \frac{(\log T)^{\frac{7}{2}}}{2^{r_{\text{last}}}} \right\} \leq P \left\{ \Delta(B) \geq \gamma \frac{2^{r_{\text{last}} + 3\beta}}{3 \times 2^{r_{\text{last}}} \min_k \beta_k(B)} \right\}
\]

\[
\leq \sum_{k \in K} P \left\{ f_k(\tilde{x}) - \frac{1}{2^{r_{\text{last}}}} \sum_{\tau = 1}^{2^{r_{\text{last}}}} y_{k,\tau}^{(B,r_{\text{last}}+1)} \geq \gamma \frac{2^{r_{\text{last}} + 3\beta}}{3 \times 2^{r_{\text{last}}} \min_k \beta_k(B)} \right\}
\]

\[
= \sum_{k \in K} P \left\{ f_k(\tilde{x}) - \eta \Delta \left( \tilde{x} ; D^{(B,r_{\text{last}}+1)}_k , 2^{-l} \right) \geq \gamma \frac{2^{r_{\text{last}} + 3\beta}}{3 \times 2^{r_{\text{last}}} \min_k \beta_k(B)} \right\}
\]

Note that since we have \( \gamma^2 \frac{2^{r_{\text{last}} + 3\beta}}{3 \times 2^{r_{\text{last}}} \min_k \beta_k(B)} \geq C_7 \), we can apply Proposition 4.5 to show that last term on the right hand side of the above display is “small.” Namely, we apply Proposition 4.5 with \( n = 2^{r_{\text{last}}} \), \( \mu = \frac{\bar{p}}{\rho_2 - d_{\text{left}}} \), \( \hat{\mu} = \frac{\bar{p}}{\rho_2 - d_{\text{left}}} \), \( \delta = \frac{\gamma_2 2^{r_{\text{last}} + 3\beta}}{3 \times 2^{r_{\text{last}}} \min_k \beta_k(B)} \), and \( h = 2^{-\beta} \) to obtain

\[
\sum_{k \in K} P \left\{ f_k(\tilde{x}) - \eta \Delta \left( \tilde{x} ; D^{(B,r_{\text{last}}+1)}_k , 2^{-l} \right) \geq \gamma \frac{2^{r_{\text{last}} + 3\beta}}{2^{r_{\text{last}}} \min_k \beta_k(B)} \right\} \leq C_7 T^{-\gamma^2 C_{18}}, \quad (A.38)
\]

where \( C_7 \) and \( C_{18} \) that depend on only \( \beta, L, \rho, \bar{p}, \) and \( d \).

### A.11 Proof of Proposition 5.5

We prove the claim by noting that if the policy reaches any bin \( B \) in the set \( \{ B : p(B) \in \tilde{\mathcal{L}}^* \} \) this implies that at least one of the bins in \( \tilde{\mathcal{L}}^* \) is replaced by its children, the probability of which can be shown to be small. On the event \( \bigcap_{B \in \mathcal{S}(B)} \), by Proposition 5.3 we need at least \( 2^{r_{\text{last}}} (\log T)^{\frac{3d+5}{2}} \times 2^{r_{\text{last}}} \min_k \beta_k(B) \) contexts to fall in \( B \in \tilde{\mathcal{L}}^* \) so that this bin is replaced by its children. By the definition of \( \tilde{\mathcal{L}}^* \), we have
\(2^{\tilde{C}} \cdot (\log T) \frac{2^{d+5}}{T} 2^d \min_k \beta_k(B) \geq T \cdot 2^{-d(B)}.\) Note that the expected number of contexts that fall in bin \(B\) over the time horizon \(T\) is \(T \cdot 2^{-d(B)}.\) Hence, by the Bernstein’s inequality in Lemma \[B.2\], we can stochastically upper bound the number of contexts that fall in a bin \(B\) at depth \(l(B):\)

\[
P\left\{ \text{\# contexts fall in } B \geq T \cdot 2^{-d(B)} \right\} \leq \exp\left( -\frac{T \cdot 2^{-d(B)}}{5} \right).
\]

Hence, by the union bound, one obtains

\[
P\left\{ \text{\# contexts fall in } B \geq T \cdot 2^{-d(B)} \text{ for at least one } B \in \tilde{\mathcal{L}}^* \right\} \leq \sum_{B \in \tilde{\mathcal{L}}^*} \exp\left( -\frac{T \cdot 2^{-d(B)}}{5} \right).
\]

Using the inequality \(e^{-x}/2 < e^{-x/2}\) for \(x \geq 0\), we have

\[
\sum_{B \in \tilde{\mathcal{L}}^*} \exp\left( -\frac{T \cdot 2^{-d(B)}}{5} \right) \leq T \frac{d}{d+2} \exp\left( -\frac{T \cdot 2^{\beta}}{5} \right),
\]

which concludes the proof. \(\Box\)

### A.12 Proof of Theorem 5.6

The proof is organized as follows. In Step 1, we lay out some notations and definitions. In Step 2, we decompose the regret with respect to non-terminal and terminal nodes. In Step 3 and 4, we analyze the regret corresponding to non-terminal nodes (the non-leaf nodes of the subtree \(\tilde{T}^*\)). Finally, in Step 5, the regret incurred on terminal nodes (leaves of the subtree \(\tilde{T}^*\)) is upper bounded.

**Step 1 (Preliminaries).** For any bin \(B \in \tilde{T}^*\), define the unique parent of \(B\) by

\[p(B) := \{ B' \in \tilde{T}^* : B \in \mathcal{C}(B) \},\]

and \(p(B) = \emptyset\) for \(B = (0,1)^d\). Moreover, let \(p^0(B) = B\) and for any \(j \geq 2\), define recursively \(p^j(B) = p(p^{j-1}(B))\). Then, the set of ancestors of any bin \(B \in \tilde{T}^*\) is denoted by \(P\) and is defined by

\[P(B) := \{ B' : B' = p^j(B) \text{ for some } j \geq 1 \} .\]
Denote by \( r_{\text{live}}(B) \) the regret incurred by the LACB policy \( \pi \) when covariate \( X_t \) fell in a live bin \( B \in L_t \), where we recall \( L_t \) is the set of live bins at time \( t \). More precisely, it is defined by

\[
r_{\text{live}}(B) := \sum_{t=1}^{T} (f^*(X_t) - f_{\pi t}(X_t)) \cdot 1 \{ B \in L_t \} \cdot 1 \{ X_t \in B \}.
\]

We also define \( B_{\text{bron}} := \bigcup_{s \leq t} L_s \) to be the set of all bins that were born at some time \( s \leq t \). We denote by \( r_{\text{born}}(B) \) the regret incurred by the policy when the covariate \( X_t \) fell in the bin \( B \). More precisely, it is defined by

\[
r_{\text{born}}(B) := \sum_{t=1}^{T} (f^*(X_t) - f_{\pi t}(X_t)) \cdot 1 \{ B \in B_{\text{bron}} \} \cdot 1 \{ X_t \in B \}.
\]

Define \( \hat{S}(B) := S(B) \cap S_{\text{last}}(B) \). Observe that Proposition 5.4 implies that if arm 1 (2) has a “larger” mean reward all over bin \( B \) then, the bin will not be eliminated and assigned to \( Q_{2,t}(Q_{1,t}) \). Hence,

\[
r_{\text{born}}(B) = r_{\text{born}}(B) \cdot 1 \{ \hat{S}(B) \} + r_{\text{live}}(B) \cdot 1 \{ \hat{S}(B) \} + \sum_{B' \in \mathcal{C}(B)} r_{\text{born}}(B') \cdot 1 \{ \hat{S}(B) \}.
\]

**Step 2 (Regret decomposition).** In what follows, we adapt the convention that \( \prod_{B' \in \mathcal{P}(B)} 1 \{ \hat{S}(B') \} = 1 \) for \( B = (0,1)^d \). The quantity we are interested in is decomposed as

\[
r(T^*) := \sum_{B \in \tilde{T}^* \setminus \tilde{L}^*} \left( r_{\text{born}}(B) \cdot 1 \{ \hat{S}(B)^c \} + r_{\text{live}}(B) \cdot 1 \{ \hat{S}(B) \} \right) \cdot \prod_{B' \in \mathcal{P}(B)} 1 \{ \hat{S}(B') \}
\]

\[
+ \sum_{B \in \tilde{L}^*} r_{\text{born}}(B) \cdot \prod_{B' \in \mathcal{P}(B)} 1 \{ \hat{S}(B') \},
\]

(A.39)

the regret accumulated on live non-terminal nodes of \( \tilde{T}^* \) and its live leaves, respectively. Our proof relies on the following events: \( \mathcal{G}(B) = \bigcup_{B' \in \mathcal{P}(B)} \hat{S}(B') \).

**Step 3 (Controlling regret on live non-terminal nodes when event \( \hat{S}(B) \) happens).** Fix a bin \( B \in \tilde{T}^* \setminus \tilde{L}^* \). On the event \( \mathcal{G}(B) \), the gap between the arms mean rewards is bounded inside \( B \) as follows:

\[
\forall x \in B : |\Delta(x)| \leq \gamma \tilde{C} (\log T)^{11/2} 2^{-l(B)} \min_k \beta_k \beta_k(B(B)).
\]

54
Furthermore, on $\hat{\mathcal{S}}^{(B)}$, the number of total samples that we collect in this bin is at most

$$6 \times 2^{r_{\text{last}}^{(B)}} \leq 642 \pi T (\log T)^{\frac{2d+5}{2}} 2^{2t(B) \min_k \beta_{f_k}(B)}.$$ 

Putting the above two displays together, one obtains

$$r_{\text{live}}^{(B)} \cdot 1 \{ \hat{\mathcal{S}}^{(B)} \} \cdot \prod_{B' \in \mathcal{P}(B)} 1 \{ \hat{\mathcal{S}}^{(B')} \} \leq \gamma C_{19} (\log T)^{\frac{2d+21}{2}} 2^{(2 \min_k \beta_{f_k}(B) - \min_k \beta_{f_k}(p(B)))},$$

for some constant $C_{19} = C_{19}(\beta, L, \bar{\rho}, \bar{\rho})$. One obtains

$$\sum_{B \in \hat{T}^* \setminus \tilde{L}^*} r_{\text{live}}^{(B)} \cdot 1 \{ \hat{\mathcal{S}}^{(B)} \} \cdot \prod_{B' \in \mathcal{P}(B)} 1 \{ \hat{\mathcal{S}}^{(B')} \} \leq \gamma C_{20} (\log T)^{\frac{2d+21}{2}} (\log T)^{\frac{10d+1}{2}} \sum_{B \in \hat{T}^* \setminus \tilde{L}^*} 2^{(2 \min_k \beta_{f_k}(B) - \min_k \beta_{f_k}(p(B)))},$$

(A.40)

**Step 4 (Controlling regret on live non-terminal nodes when event $\hat{\mathcal{S}}^{(B)}$ does not happen).**

By Propositions 5.3 and 5.4, one has

$$\mathbb{E} \left[ \sum_{B \in \hat{T}^* \setminus \tilde{L}^*} r_{\text{born}}^{(B)} \cdot 1 \{ \hat{\mathcal{S}}^{(B)} \} \cdot \prod_{B' \in \mathcal{P}(B)} 1 \{ \hat{\mathcal{S}}^{(B')} \} \right] \leq C_{21} (\log T)^{\frac{2d}{2}} T^{-\gamma C_{22} + 1} \sum_{l=0}^{\left\lfloor \frac{1}{\sqrt{T}} \log T \right\rfloor} 2^{ld} \leq 4C_{21} (\log T)^{\frac{2d}{2}} T^{-\gamma C_{22} + 1 + \frac{d+2}{5}},$$

(A.41)

for some constants $C_{21} = C_{21}(\beta, L, \bar{\rho}, \bar{\rho}, d)$ and $C_{22} = C_{22}(\beta, L, \bar{\rho}, \bar{\rho}, d)$.

**Step 5 (Controlling regret on live terminal nodes).** Note that by Proposition 5.5, we have

$$\mathbb{E} \left[ \sum_{B \in \tilde{L}^*} r_{\text{born}}^{(B)} \left\{ \bigcap_{B \in \tilde{L}^*} \mathcal{S}^{(B)} \right\} \right] \leq T \cdot T^{\frac{d}{2+2d}} \exp \left( -T^{\frac{2d}{2+2d}} / 5 \right),$$

which implies

$$\mathbb{E} \left[ \sum_{B \in \tilde{L}^*} r_{\text{born}}^{(B)} \right] \leq T \cdot T^{\frac{d}{2+2d}} \exp \left( -T^{\frac{2d}{2+2d}} / 5 \right) + T \cdot \mathbb{P} \left\{ \bigcup_{B \in \tilde{L}^*} \mathcal{S}^{(B)} \right\} \leq T \cdot T^{\frac{d}{2+2d}} \exp \left( -T^{\frac{2d}{2+2d}} / 5 \right) + C_{21} (\log T)^{\frac{2d}{2}} T^{-\gamma C_{22} + 1 + \frac{d}{2+2d}},$$

(A.42)
where the second inequality follows from Propositions 5.3 and 5.4. Putting together (A.39), (A.40), (A.41), and (A.42), the desired result follows. \qed
B Auxiliary lemmas

Lemma B.1 (Chernoff-Hoeffding bound). Let $X_1, \ldots, X_n$ be random variables such that $X_t$ is a $\sigma^2$-sub-Gaussian random variable conditioned on $X_1, \ldots, X_{t-1}$ and $\mathbb{E}[X_t | X_1, \ldots, X_{t-1}] = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$

$$
\mathbb{P}\{S_n \geq n\mu + a\} \leq e^{-\frac{a^2}{2n\sigma^2}}, \quad \text{and} \quad \mathbb{P}\{S_n \leq n\mu - a\} \leq e^{-\frac{a^2}{2n\sigma^2}}.
$$

Lemma B.2 (Bernstein inequality). Let $X_1, \ldots, X_n$ be random variables with range $|X_t| \leq M$ and

$$
\sum_{t=1}^{n} \text{Var}[X_t | X_{t-1}, \ldots, X_1] = \sigma^2.
$$

Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$

$$
\mathbb{P}\{S_n \geq \mathbb{E}[S_n] + a\} \leq \exp \left( -\frac{a^2/2}{\sigma^2 + Ma/3} \right).
$$

References


Qiang, S. and M. Bayati (2016). Dynamic pricing with demand covariates. *Available at SSRN 2765257*.


