Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Smooth Contextual Bandits: Bridging the Parametric and Nondifferentiable Regret Regimes

Yichun Hu, Nathan Kallus, Xiaojie Mao

To cite this article:
Yichun Hu, Nathan Kallus, Xiaojie Mao (2022) Smooth Contextual Bandits: Bridging the Parametric and Nondifferentiable Regret Regimes. Operations Research

Published online in Articles in Advance 07 Feb 2022

. https://doi.org/10.1287/opre.2021.2237

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–21
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Smooth Contextual Bandits: Bridging the Parametric and Nondifferentiable Regret Regimes

Yichun Hu,[a] Nathan Kallus,[a,*] Xiaojie Mao[b]

[a] Cornell University, New York, New York 10044; [b] Tsinghua University, 100084 Beijing, China
*Corresponding author

Contact: yh767@cornell.edu, https://orcid.org/0000-0002-5826-9665 (YH); kallus@cornell.edu, https://orcid.org/0000-0002-2757-1570 (NK); maoxj@sem.tsinghua.edu.cn, https://orcid.org/0000-0003-2985-1741 (XM)

Received: September 6, 2020
Revised: April 26, 2021; August 23, 2021
Accepted: October 14, 2021
Published Online in Articles in Advance: February 7, 2022

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2021.2237

Abstract. We study a nonparametric contextual bandit problem in which the expected reward functions belong to a Hölder class with smoothness parameter β. We show how this interpolates between two extremes that were previously studied in isolation: nondifferentiable bandits (β at most 1), with which rate-optimal regret is achieved by running separate noncontextual bandits in different context regions, and parametric-response bandits (infinite β), with which rate-optimal regret can be achieved with minimal or no exploration because of infinite extrapolatability. We develop a novel algorithm that carefully adjusts to all smoothness settings, and we prove its regret is rate-optimal by establishing matching upper and lower bounds, recovering the existing results at the two extremes. In this sense, our work bridges the gap between the existing literature on parametric and nondifferentiable contextual bandit problems and between bandit algorithms that exclusively use global or local information, shedding light on the crucial interplay of complexity and regret in contextual bandits.

Funding: This work was supported by the National Science Foundation Division of Information and Intelligent Systems [Grant 1846210].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2021.2237.

Keywords: contextual bandits • local polynomial regression • minimax regret • margin condition

1. Introduction

In many domains, including healthcare and e-commerce, we frequently encounter the following decision-making problem: we sequentially and repeatedly receive context information X (e.g., features of patients or users), need to choose an action $A \in \mathcal{A}$ from among $|\mathcal{A}| < \infty$ actions (e.g., with which therapy, if any, to treat a patient or which ad, if any, to show to a user), and receive a reward $Y(A)$ (e.g., patient's health outcome or user's click minus ad spot costs) corresponding to the chosen action. Our goal is to collect the most reward over time. When contexts X and potential rewards $\{Y(a) : a \in \mathcal{A}\}$ are drawn from a stationary but unknown distribution, this setting is modeled by the stochastic bandit problem (Wang et al. 2005, Bubeck and Cesa-Bianchi 2012). A special case is the multiarmed bandit (MAB) problem in which there is no contextual information (Lai and Robbins 1985, Auer et al. 2002). In these problems, we quantify the quality of an algorithm for choosing actions based on available historical data in terms of its *regret* for every horizon T: the expected additional cumulative reward up to time T that we would obtain if we had full knowledge of the stationary context–reward distribution (but not the realizations). The *minimax regret* is the best (over algorithms) worst-case regret (over problem instances).

The relevant part of the context–reward distribution for maximum expected reward decision making is the conditional mean reward functions, $\eta_a(x) = \mathbb{E}[Y(a)|X = x]$ for $a \in \mathcal{A}$: if we knew these functions, we would know what arm to pull. Because we only observe the reward of the chosen action, $Y(A)$, and never that of the unchosen actions, $Y(a)$ $\forall a \neq A$, we face the oft-noted trade-off between exploration and exploitation: we are motivated to greedily exploit the arm we currently think is best for the context so as to collect the highest reward right now, but we also need to explore other arms to learn about its expected reward function for fear of missing better options in the future because of lack of information.

The trade-off between exploration and exploitation crucially depends on how we model the relationship between the context and the reward, that is, η_a. When

we restrict η_a to a *model*, such as linear functions, minimax regret gives rigorous meaning to our not knowing the particular instance being faced at the onset and needing to *learn* the reward structure. Specifically, it answers the question, given only the information that η_a belongs to a certain model, how small can one ensure regret is no matter what by learning and adapting to any one instance. In the stochastic setting, previous literature considers two extreme cases in isolation: a parametric reward model, usually linear (Goldenshluger and Zeevi 2013, Bastani and Bayati 2020, Bastani et al. 2020), and a nonparametric, nondifferentiable reward model (Rigollet and Zeevi 2010, Perchet and Rigollet 2013, Fontaine et al. 2019). We review these before describing our contribution. We define the problem in complete formality in Section 2.

Linear-Response Bandit

One extreme is the linear-response bandit with which the expected reward function is assumed to be linear in context, $\eta_a(x) = \theta_a^\top x$ (Goldenshluger and Zeevi 2013, Bastani and Bayati 2020). This parametric assumption imposes a global structure on the expected reward function and permits extrapolation because all samples from arm a are informative about the finite-dimensional parameters θ_a regardless of the context (see Figure 1(a)). Dramatically, this global structure almost entirely obviates the need for forced exploration. In particular, Bastani et al. (2020) prove that, under very mild conditions, the greedy algorithm is rate-optimal for linear reward models, achieving logarithmic regret. Consequently, the result shows that the classic trade-off that characterizes contextual bandit problems is often not present in linear-response bandits. Similar behavior generally occurs when we impose other parametric models on expected rewards. At the same time, whereas theoretically, regret is consequently very low, linear- and parametric-response bandit algorithms may actually have linear regret in practice because the parametric assumption usually fails to hold exactly.

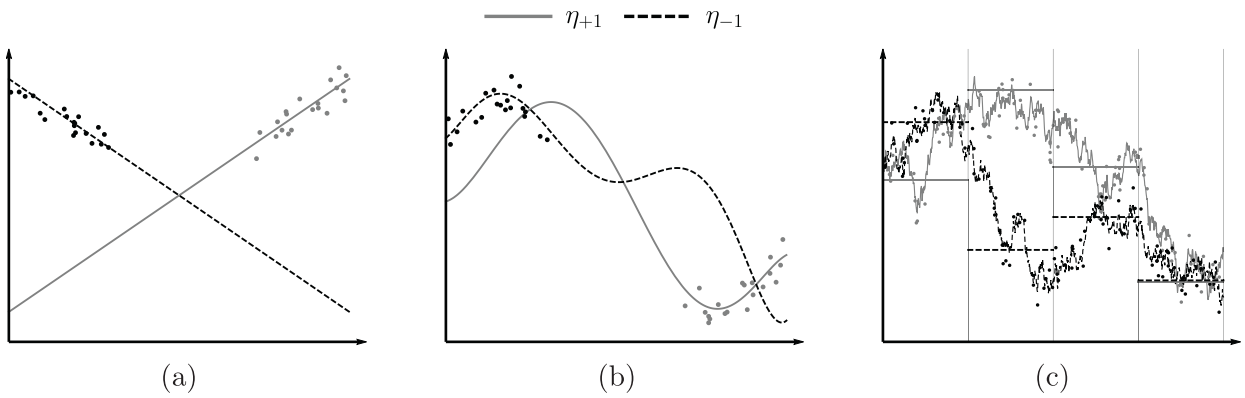
Nondifferentiable Nonparametric-Response Bandit

Another line of literature considers nonparametric reward models that satisfy a Hölder continuity condition (Rigollet and Zeevi 2010, Perchet and Rigollet 2013), the strongest form of which is Lipschitz continuity. In stark contrast to the linear case, such functions need not even be differentiable. (Note the difference to Hölder *smoothness*, which imposes Hölder continuity on *derivatives*.) In any nonparametric-response bandit, extrapolation is limited because only nearby samples are informative about the reward functions at each context value (Figure 1(b)). Thus, we need to take a more localized learning strategy: we have to actively explore in *every* context region and learn the expected reward functions using nearby samples. In the nondifferentiable extreme, Rigollet and Zeevi (2010) show that one can achieve rate-optimal regret by partitioning the context space into small hypercubes and running completely separate MAB algorithms (e.g., upper confidence bound (UCB)) within each hypercube in isolation (Figure 1(c)). In other words, we can almost ignore the contextual structure because we obtain so little information across contexts. However, the regret is also correspondingly very high.

Our Contribution: Smooth Contextual Bandits

In this paper, we consider a nonparametric-response bandit problem with smooth expected reward functions. This bridges the gap between the infinitely smooth linear-response bandit and the unsmooth nondifferentiable-response bandit. We characterize the

Figure 1. The Fundamental Nature of Contextual Bandit Problems Depends Crucially on the Assumed Structure of Expected Reward Functions, η_a for Two Arms $a \in A = \{-1, +1\}$



Notes. (a) A linear response bandit: samples in one context region are fully informative about expected rewards in any other context region. (b) A nonparametric-response bandit: samples offer only limited extrapolation to learn expected rewards at nearby context values. (c) A nondifferentiable-response bandit: rate-optimal regret obtainable by reducing the contextual bandit into multiple, separate MAB problems.

smoothness of the expected reward functions in terms of the highest order of continuous derivatives or, more generally, in terms of a Hölder smoothness parameter β, which generalizes both nondifferentiable Hölder continuous functions ($\beta \le 1$) and infinitely extrapolatable functions (such as linear, which we denote by $\beta = \infty$). Table 1 summarizes the landscape of the current literature and where our paper lies in terms of this new smoothness perspective and the sharpness α of the margin (see Assumption 4).

We propose a novel algorithm for every level of smoothness $1 \le \beta < \infty$ and prove that it achieves the minimax optimal regret rate up to polylogs. In particular, when $\beta > 1$, we must leverage information across farther apart contexts, and running separate MAB algorithms is suboptimal. And, because $\beta < \infty$, we must ensure sufficient exploration everywhere. Thus, our algorithm interpolates between the fully global learning of the linear-response bandit (which satisfies $\beta = \infty$) and the fully local learning of the nondifferentiable bandit ($0 < \beta \le 1$), according to the smoothness of the expected reward functions. The smoother the expected reward functions, the more global reward information we incorporate. Moreover, our algorithm judiciously balances exploration and exploitation: it exploits only when we have certainty about which arm is optimal, and it explores economically in a shrinking margin region with fast diminishing error costs. As a result, our algorithm achieves regret bounded by $\tilde{O}\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}+1\right)$ (where $\tilde{O}$ means up to polylogarithmic factors). We show that, for any algorithm, there exists an instance on which it must have regret lower bounded by the same rate, showing that our algorithm is rate-optimal and establishing the minimax regret rate for the problem. Consequently, the minimax regret, $\mathcal{R}_T$, which we define in Section 2.6, satisfies $\mathcal{R}_T = \tilde{\Theta}\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}+1\right)$, and hence, $\lim_{T\to\infty} \log(\mathcal{R}_T)/\log(T) = \frac{(\beta+d-\alpha\beta)_+}{2\beta+d}$, where $(\cdot)_+$ denotes $\max\{\cdot, 0\}$.

Whereas this rate has the same *form* as the regret in the nondifferentiable case studied by Rigollet and Zeevi (2010), our results extend to the smooth ($\beta > 1$) regime in which our algorithm can attain much lower regret, arbitrarily approaching polylogarithmic rates as smoothness increases. Our algorithm is fundamentally different, leveraging contextual information from farther away as smoothness increases without deteriorating estimation resolution, and our analysis is necessarily much finer. Our work connects seemingly disparate contextual bandit problems and reveals the whole spectrum of minimax regret over varying levels of function complexity.

1.1. Related Literature

Nonparametric Regression. Our algorithm leverages nonparametric regression to learn expected reward functions, namely, local polynomial regression. Nonparametric regression seeks to estimate regression (aka conditional expectation) functions without assuming that they belong to an a priori known parametric family. One of the most popular nonparametric regression methods is the Nadaraya–Watson kernel regression estimator (Nadaraya 1964, Watson 1964), which estimates the conditional expectation at a query point as the weighted average of observed outcomes, weighted by their closeness to the query using a similarity-measuring function known as a kernel. Local polynomial estimators generalize this by fitting a polynomial by kernel-weighted least squares (Stone 1977), in which fitting a constant recovers the former. Stone (1980) considers function classes with different levels of smoothness β and shows that local polynomial regression achieves rate-optimal point convergence. Stone (1982) further shows that a modification of this estimator can achieve rate-optimal convergence in p-norm for $0 < p \le \infty$. There are a variety of other nonparametric estimators that can achieve rate optimality in these classes, such as sieve estimators (e.g., Chen 2007, Belloni et al. 2015), but we do not use these in our algorithm. For more detail and an exhaustive bibliography on nonparametric regression, see Tsybakov (2008).

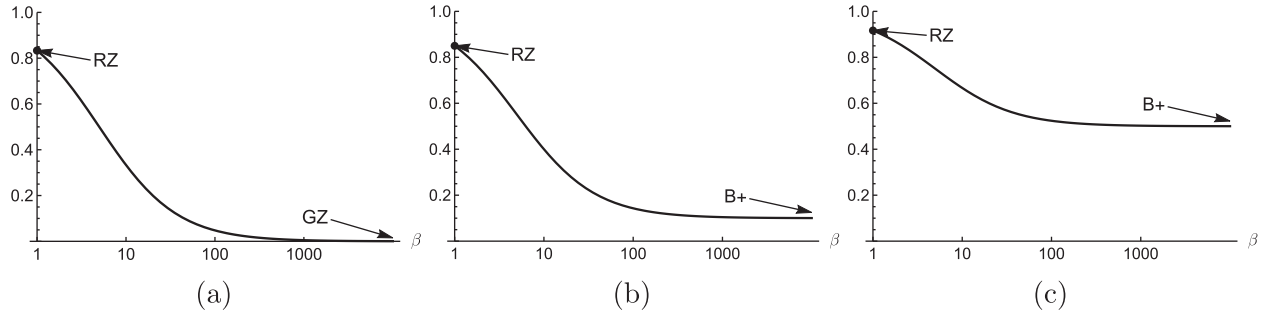
Nonparametric regression also has broad applications in decision making. In classification problems, Audibert and Tsybakov (2007) establish fast convergence rates for the zero–one error of plug-in estimators based on local polynomial regression by leveraging a finite-

Table 1. The Lay of the Literature on Stochastic Contextual Bandits in Terms of Our Smoothness Perspective

		$\beta \le 1$	***Smoothness*** $1 \le \beta < \infty$	$\beta = \infty$
Margin sharpness	$0 \le \alpha < 1$	Rigollet and Zeevi (2010)	*This paper*	Bastani et al. (2020)
	$\alpha = 1$			Goldenshluger and Zeevi (2013)
	$\alpha > 1$	Perchet and Rigollet (2013)		Bastani et al. (2020)

Notes. For the most part, there has been a significant and wide divide between nondifferentiable- and parametric-response bandits. Our work shows that (up to polylogs) the minimax regret rate $\tilde{\Theta}(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}+1)$ reigns across *all* regimes; see also Figure 2. (Note that additional linear restrictions are made in the $\beta = \infty$ column.)

Figure 2. The Minimax Regret Rate Exponent, $\lim_{T\to\infty}\log(\mathcal{R}_T)/\log(T) = \frac{(\beta+d-\alpha\beta)_+}{2\beta+d}$, as Shown by Our Theorems 2 and 3



Notes. The minimax regret $\mathcal{R}_T$ is defined in Section 2.6. Existing results shown with arrows only characterize the two extreme regimes (RZ refers to Rigollet and Zeevi 2010, GZ refers to Goldenshluger and Zeevi 2013, B+ refers to Bastani et al. 2020). In between, our results reveal the effect of complexity on regret. (a) $\alpha = 1, d = 10$. (b) $\alpha = 0.8, d = 10$. (c) $\alpha = 0, d = 10$.

sample concentration bound. The rate depends on a so-called margin condition number α originally proposed by Mammen and Tsybakov (1999) and Tsybakov (2004) that quantifies how well separated the classes are when larger α corresponds to more separation (see Assumption 4). Bertsimas and Kallus (2019) use similar locally weighted nonparametric regression methods to solve conditional stochastic optimization problems with auxiliary observations and show that this provides model-free asymptotic optimality.

Contextual Bandits. Whereas the literature noted usually considers an off-line problem with a given exogenous sample of data, the literature on contextual bandit problems considers adaptive data collection and sequential decision making (see Bubeck and Cesa-Bianchi 2012 for a complete bibliography). Some contextual bandit literature allows for adversarially chosen contexts (e.g., Langford and Zhang 2007, Beygelzimer et al. 2011), but this leads to high regret and may be too pessimistic in real-world applications. For example, in clinical trials for a noninfectious disease, the treatment decisions for one patient do not have direct impacts on the personal features of the next patient. One line of literature captures this stochastic structure by assuming that contexts and rewards are drawn independently and identically distributed (i.i.d.) from a stationary but unknown distribution (e.g., Wang et al. 2005, Dudik et al. 2011, Agarwal et al. 2014). The aforementioned linear- and nonparametric-response bandits both fall into this setting. Rigollet and Zeevi (2010), Goldenshluger and Zeevi (2009, 2013), and Perchet and Rigollet (2013) introduce the use of the margin condition in this setting to quantify how well separated the arms are, a well-known determiner of regret in the simpler MAB problem (Lai and Robbins 1985).

Goldenshluger and Zeevi (2013) assume a linear model between rewards and covariates for each arm and propose a novel rate-optimal algorithm that works by maintaining two sets of parameter estimates for each arm. Bastani et al. (2020) show that the greedy algorithm is optimal under mild covariate diversity conditions. Bastani and Bayati (2020) consider a sparse linear model and use a LASSO estimator to accommodate high-dimensional contextual features. Whereas Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020) assume a sharp margin ($\alpha = 1$), Goldenshluger and Zeevi (2009) also consider more general margin conditions in the one-armed linear-response setting and (Bastani et al. 2020, appendix E) considers these in the multiarmed linear-response setting. All of these achieve regret bounds of order $\log T$ under a sharp margin condition ($\alpha = 1$). However, as discussed, this relies heavily on the fact that every observation is informative about expected rewards everywhere.

Valko et al. (2013) assume that arm rewards belong to a reproducing kernel Hilbert space (RKHS) with a bounded kernel function (e.g., Gaussian). Whereas this model considerably generalizes the linear model, it is similar in two crucial ways: the learning rate is similar and extrapolation is still possible. For off-line regression in an RKHS, the rate is at worst $O(n^{-1/2})$ and at best $\tilde{O}(n^{-1})$ (Bartlett et al. 2005, corollary 6.7), which stands in stark contrast to the rate possible when only assuming limited differentiability, which only approaches $O(n^{-1/2})$ as the number of derivatives increases infinitely (Stone 1980, 1982). Furthermore, assuming a bounded RKHS norm essentially enables extrapolation: for example, for the Gaussian kernel, if two functions agree on a nonempty open set, they agree everywhere, meaning we can extrapolate from such a subset (Steinwart et al. 2006, corollary 3.9). In contrast, the lower bound we prove on regret in our problem (Theorem 3) relies on constructing an example with arbitrary constant values in different regions, forcing one to explore each region as extrapolation is not possible. Valko et al. (2013) indeed obtain a regret bound of $\tilde{O}(\sqrt{T})$, which matches the bounds for linear response (or our bound as $\beta \to \infty$) without a margin condition ($\alpha = 0$) as Valko et al. (2013) indeed do not impose the margin condition.

Rigollet and Zeevi (2010) and Perchet and Rigollet (2013) study the case in which we only assume that the expected reward functions are Hölder continuous, that is, that $|\eta_a(x) - \eta_a(x')| \le \|x - x'\|^\beta$. Note that $\beta = 1$ corresponds to Lipschitz continuity and is the strongest variant of this assumption because $\beta > 1$ requires the function to be constant and is, therefore, not considered. Rigollet and Zeevi (2010) study the two-arm case and obtain optimal minimax-regret rates for margin condition $\alpha \le 1$. The rate optimal algorithm in this case (UCBogram) consists of segmenting the context space at the beginning and running separate MAB algorithms in parallel in each segment. Perchet and Rigollet (2013) extend this to multiple arms and any $\alpha \ge 0$ by proposing another algorithm (adaptively binned successive elimination (ABSE)) that gradually refines the segmentations of the context space (hence, avoiding pulling each arm in each of very many segments when the arm separation is strong) but still only uses data *within* each segment to estimate the reward functions in that segment. Crucially, this hyperlocal approach is no longer rate-optimal when we impose smoothness, with which we must use information from *across* such segments to fully leverage reward smoothness.

Reeve et al. (2018) also consider Lipschitz expected reward functions ($\beta = 1$) but leverage a k-nearest neighbor regression algorithm in order to adapt to the underlying dimension of the support of covariates. Their algorithm also avoids the need to segment the context space. The regret bound is the same as Rigollet and Zeevi (2010) and Perchet and Rigollet (2013) with d replaced by the underlying dimension, which may be smaller than the ambient dimension. In particular, whereas they can leverage the lower underlying dimension when it exists, they cannot leverage higher order differentiability. Slivkins (2011) considers a possibly infinite number of arms and assumes $\eta_a(x)$ is *jointly* Lipschitz in (a, x). When the number of arms is finite, the regret bound matches Rigollet and Zeevi (2010) and Perchet and Rigollet (2013) (or our bound with $\beta = 1$) without margin a condition ($\alpha = 0$), which Slivkins (2011) does not impose.

As in Goldenshluger and Zeevi (2013), Rigollet and Zeevi (2010), and Perchet and Rigollet (2013), our work focuses on computing the minimax regret rate, which is defined for a given class of bandit problem instances. And, as in Rigollet and Zeevi (2010) and Perchet and Rigollet (2013), the class of instances we consider is parameterized by a constant β controlling the smoothness of expected reward functions, and we compute the minimax regret rate for *each* β. The minimax regret is defined as the infimum over policies of the supremum over instances in the class (see Section 2.6). Although the infimizer (over policies) in the minimax regret cannot know the instance chosen by the inner supremum, it does know the class of instances available to it. Both the previously cited works and our work, therefore, compute an upper bound on the minimax regret by exhibiting a policy that *depends* on the class of instances being considered in the supremum and, therefore, on β in our case.

In addition to computing the minimax regret for each β, an important supplementary question is *adaptability* to β: does there exist a policy that does not depend on β yet achieves the minimax regret rate for each β? This question depends, of course, on first computing the minimax regret rate for each β. Since our paper and based on our work, Gur et al. (2019) answered this question negatively in general and positively if one further assumes a self-similarity condition on expected reward functions. Under this assumption, they show that, for adaptation, it suffices to first explore arms evenly for $o\left(T^{1-\frac{\beta(\alpha+1)}{2\beta+d}}\right)$ time, which only adds to the regret terms that are lower order than what we show is the optimal rate, then use the collected data to estimate β by $\hat{\beta}$, and then run our nonadaptive algorithm (Algorithm 1) with the smoothness parameter set to $\hat{\beta}$ if $\hat{\beta} > 1$ or run the nonadaptive algorithm of Perchet and Rigollet (2013) with the smoothness parameter set to $\hat{\beta}$ if $\hat{\beta} \le 1$.

1.2. Notation

For any multiple index $r = (r_1, \dots, r_d) \in \mathbb{Z}_+^d$ and any $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, define $|r| = \sum_{i=1}^d r_i$, $r! = r_1! \cdots r_d!$, $x^r = x_1^{r_1} \cdots x_d^{r_d}$ and the differential operator $D^r := \frac{\partial^{r_1+\cdots r_d}}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}}$. We use $\|\cdot\|$ to represent the Euclidean norm and Leb$[\cdot]$ the Lebesgue measure. We let $\mathcal{B}(x,h) = \{x' \in \mathbb{R}^d : \|x' - x\| \le h\}$ be the ball with center x and radius $h > 0$, and $v_d = \pi^{d/2}/\Gamma(d/2+1)$, the volume of a unit ball in $\mathbb{R}^d$. For any $\beta > 0$, let $\mathfrak{b}(\beta) = \sup\{i \in \mathbb{Z} : i < \beta\}$ be the maximal integer that is *strictly* less than β, and let M_β be the cardinality of the set $\{r \in \mathbb{Z}_+^d : |r| \le \mathfrak{b}(\beta)\}$. For an event A, the indicator function $\mathbb{I}(A)$ is equal to one if A is true and zero otherwise. For two scalars $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For a matrix $\mathcal{A}$, its minimum eigenvalue is denoted as $\lambda_{\min}(\mathcal{A})$. For two functions $f_1(T) > 0$ and $f_2(T) > 0$, we use the standard notation for asymptotic order: $f_1(T) = O(f_2(T))$ represents $\limsup_{T\to\infty} \frac{f_1(T)}{f_2(T)} < \infty$, $f_1(T) = \Omega(f_2(T))$ represents $\liminf_{T\to\infty} \frac{f_1(T)}{f_2(T)} > 0$, and $f_1(T) = \Theta(f_2(T))$ represents simultaneously $f_1(T) = \Omega(f_2(T))$ and $f_1(T) = O(f_2(T))$. We use $\tilde{O}, \tilde{\Omega}, \tilde{\Theta}$ to represent the same order relationship up to polylogarithmic factors. For example, $f_1(T) = \tilde{O}(f_2(T))$ means $\limsup_{T\to\infty} \frac{f_1(T)}{\text{polylog}(T) f_2(T)} < \infty$ for a polylogarithmic function polylog(T).

1.3. Organization

The rest of the paper is organized as follows. In Section 2, we formally introduce the smooth nonparametric bandit problem and assumptions. For a lucid exposition, we focus on the setting of two arms in the main text and study the more general multiarmed setting in Online Appendix A. We describe our proposed algorithm in Section 3. In Section 4, we analyze our algorithm theoretically: we derive an upper bound on the regret of our algorithm in Section 4.1, and we prove a matching lower bound on the regret of any algorithm in Section 4.2. We provide a numerical investigation with a simplified version of our algorithm in Section 5. We conclude our paper in Section 6. Whereas proof techniques are outlined, complete proof details are relegated to the online appendix.

2. The Smooth Contextual Bandit Problem

In this section, we formulate the smooth contextual bandit problem that we consider in this paper. We break up this formulation into parts, explaining the significance or necessity of each part separately. We focus on the two-armed smooth contextual bandit problem, letting $\mathcal{A} = \{-1, +1\}$. We extend the problem, our algorithm, and our analysis to multiarmed problems in Online Appendix A.

2.1. Two-Armed Stochastic Contextual Bandits

Consider the following two-armed contextual bandit problem. For $t = 1, 2, \ldots$, nature draws $(X_t, Y_t(1), Y_t(-1))$ i.i.d. from a common distribution $\mathbb{P}$ of $(X, Y(1), Y(-1))$, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is the context (covariate) and $Y(\pm 1) \in [0,1]$ are random rewards corresponding to arm ± 1. At each time step t, the decision maker observes the context X_t, pulls an arm $A_t \in \{-1, 1\}$ according to the observed context and history so far, and then obtains the reward $Y_t = Y_t(A_t)$ of the chosen arm. Specifically, an admissible policy (allocation rule), $\pi = \{\pi_t\}$, is a sequence of *random* functions $\pi_t : \mathcal{X} \to \{-1, 1\}$ such that, for each t, π_t is conditionally independent of $(X_1, A_1, Y_1(1), Y_1(-1), \ldots)$ given $(X_1, A_1, Y_1, \ldots, X_{t-1}, A_{t-1}, Y_{t-1})$, and we let $A_t = \pi_t(X_t)$, $Y_t = Y_t(A_t)$.

For $x \in \mathcal{X}$, we denote the conditional expected reward functions as

$$\eta_{\pm 1}(x) = \mathbb{E}[Y(\pm 1) | X = x],$$

and the conditional average treatment effect (CATE) of pulling arm 1 versus arm –1 as

$$\tau(x) = \mathbb{E}[Y(1) | X = x] - \mathbb{E}[Y(-1) | X = x] = \eta_1(x) - \eta_{-1}(x).$$

Obviously, if we had full knowledge of the regression functions $\eta_{\pm 1}$ or the CATE function τ, the optimal decision at each time step would be the oracle policy π^* that always pulls the arm with higher expected reward given X_t and regardless of history, namely,

$$\pi^*(x) = \mathbb{I}(\tau(x) \geq 0) - \mathbb{I}(\tau(x) < 0) \in \operatorname{argmax}_{a \in \{-1,1\}} \eta_a(x). \tag{1}$$

However, because we do not know these functions, the oracle policy is infeasible in practice. We measure the performance of a policy π by its *(expected cumulative) regret* compared with the oracle policy π^* up to any time T, which quantifies how much the policy π is inferior to the oracle policy π^*:

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} (Y_t(\pi^*(X_t)) - Y_t(\pi_t(X_t)))\right]. \tag{2}$$

The growth of this function in T quantifies the quality of π.

2.2. Smooth Rewards

In this paper, we aim to construct a decision policy that achieves low regret *without* strong parametric assumptions on the expected reward functions. We instead focus on expected reward functions restricted to a Hölder class of functions. This is the key restriction characterizing the nature of the bandit problem we consider.

Definition 1 (Hölder Class of Functions). A function $\eta : \mathcal{X} \to [0,1]$ belongs to the $(\beta, L, \mathcal{X})$-Hölder class of functions if it is $\mathfrak{b}(\beta)$-times continuously differentiable and, for any $x, x' \in \mathcal{X}$,

$$\left| \eta(x') - \sum_{|r| \leq \mathfrak{b}(\beta)} \frac{(x' - x)^r}{r!} D^r \eta(x) \right| \leq L \|x' - x\|^\beta. \tag{3}$$

Recall that $\mathfrak{b}(\beta)$ is the largest integer strictly smaller than β. When $\beta \leq 1$, Equation (3) reduces to Hölder continuity (i.e., $|\eta(x) - \eta(x')| \leq L\|x' - x\|^\beta$), as considered in previous nondifferentiable bandit literature (Rigollet and Zeevi 2010, Perchet and Rigollet 2013). When $\beta > 1$, $\mathfrak{b}(\beta)$ is the highest order of continuous derivatives. For example, when $\mathcal{X}$ is compact, k-times continuously differentiable functions are $(k, L, \mathcal{X})$-Hölder for some L. Polynomials of bounded degree k are $(\beta, 0, \mathcal{X})$-Hölder for all $\beta > k$.

In this paper, we focus on $\beta \geq 1$, which crucially includes the smooth case ($\beta > 1$).

Assumption 1 (Smooth Conditional Expected Rewards). *For $a = \pm 1$, η_a is $(\beta, L, \mathcal{X})$-Hölder for $\beta \geq 1$ and is also $(1, L_1, \mathcal{X})$-Hölder.*

Given a function that is $(\beta, L, \mathcal{X})$-Hölder on a compact $\mathcal{X}$ with $\beta \geq 1$, there *always* exists a finite $L_1 > 0$ such that the function is also $(1, L_1, \mathcal{X})$-Hölder (i.e., L_1-Lipschitz). Thus, assuming Lipschitzness in the second part of Assumption 1 is actually not necessary for

characterizing the regret rate of our algorithm for any single, fixed instance if we assume a compact $\mathcal{X}$ as we do in Assumption 3. However, from the perspective of characterizing the minimax regret, with which we take a supremum over instances, it is necessary as the Lipschitz constant L_1 may be arbitrarily large in the $(\beta, L, \mathcal{X})$-Hölder class of functions.

2.3. Optimal Decision Region Regularity

We next introduce a regularity condition on the context regions in which each arm is optimal, namely,

$$\mathcal{Q}_a = \{x \in \mathcal{X} : a\tau(x) \geq 0\}.$$

When the expected rewards are not restricted parametrically as we imposed earlier, we *must* use local information to estimate them because extrapolation is limited. In particular, in order to estimate $\eta_a(x)$ consistently at a given point x, we must have that the contexts of our data on outcomes from arm a eventually become dense around the point x. To formalize this notion, we introduce weak and strong (c_0, r_0)-regularity conditions.

Definition 2 ((c_0, r_0)-Regularity). A Lebesgue-measurable set $\mathcal{S} \subseteq \mathbb{R}^d$ is called weakly (c_0, r)-regular at point $x \in \mathcal{S}$ if

$$\text{Leb}[\mathcal{S} \cap \mathcal{B}(x, r)] \geq c_0 \text{Leb}[\mathcal{B}(x, r)].$$

If this condition holds for all $0 \leq r \leq r_0$, then set $\mathcal{S}$ is called strongly (c_0, r_0)-regular at x. Furthermore, if $\mathcal{S}$ is strongly (c_0, r_0)-regular at all $x \in \mathcal{S}$, then the set $\mathcal{S}$ is called a strongly (c_0, r_0)-regular set.

Essentially, if our data for arm a became dense in the set $\mathcal{S}$ and if $\mathcal{S}$ is strongly (c_0, r_0)-regular at x, then sufficient data are available within any small neighborhood around x to estimate $\eta_a(x)$ well. If $\mathcal{S}$ is not strongly regular, then even if our data became dense in $\mathcal{S}$, there would be diminishing amounts of data available as we looked closer and closer near x. For example, the ℓ_q unit ball is strongly regular for $q \geq 1$ and irregular for $q < 1$ because the points at its corners are too isolated from the rest of the set.

Naturally, we need enough data from arm a around x to estimate $\eta_a(x)$ accurately. Luckily, we need only worry about high-accuracy estimation for *both* arms near the decision boundary, at which it is hard to tell which of the arms is optimal. (Intuitively, away from the boundary, it is very easy to separate the arms with very few samples as in the classic MAB case of Lai and Robbins 1985.) But we cannot rely on having enough data from arm a in a whole ball around every point near the boundary as that would require us to pull arm a too often across the boundary, in $\mathcal{Q}_{-a}$, where it is not optimal. This would necessarily lead to high regret. Instead, we must be able to rely mostly on data from arm-a pulls in $\mathcal{Q}_a$. Therefore, we must have that this set is strongly regular. If, otherwise, there exists such a point $x \in \mathcal{Q}_a$ that is sufficiently isolated from the rest of $\mathcal{Q}_a$, then we cannot generate enough samples to learn $\eta_a(x)$ accurately enough without necessarily incurring high regret.

Assumption 2 (Optimal Decision Regions). *For $a = \pm 1$, $\mathcal{Q}_a$ is a nonempty strongly (c_0, r_0)-regular set.*

An illustration of this condition is given in Figure 3. We note that this condition is a refinement of the usual condition for nonparametric estimation, which simply requires the support $\mathcal{X}$ to be a strongly regular set (Tsybakov 2008). This refinement is necessary for the unique bandit setting we consider in which we must worry about the costs of adaptive data collection and may not simply assume a good data set is given. Because the intersection of strongly regular sets may not always be strongly regular, it is insufficient to only assume the support $\mathcal{X}$ is strongly regular and expected rewards are smooth in order to guarantee Assumption 2, as seen in Figure 3(b).

Assumption 2 is *necessary* to guarantee the optimal minimax regret regime we study, bridging the previous nondifferentiable and parametric regimes. In particular, we show that, for any policy, there exist problem instances satisfying all assumptions except Assumption 2 for which the regret rate is *higher* than the minimax regret rate for instances satisfying all assumptions (see Theorem 4).

We end this section by demonstrating the value of c_0 in concrete examples. If $\mathcal{Q}_a = [0, r_0]^d$ is a d-dimensional cube, we can easily verify that it is strongly $(2^{-d}, r_0)$-regular. If $\mathcal{Q}_a = \mathcal{B}((\frac{1}{2}, \ldots, \frac{1}{2}), r_0)$ is a d-dimensional ball, we can show that it is strongly $\left(\frac{(\sqrt{3})^{d-1} v_{d-1}}{2^{d-2} d v_d}, r_0\right)$-regular. Similarly, if $\mathcal{Q}_a = \mathcal{B}((\frac{1}{2}, \ldots, \frac{1}{2}), r_0) \cap \{x : x_1 \geq \frac{1}{2}\}$ is a half ball, then it is strongly $\left(\frac{(\sqrt{3})^{d-1} v_{d-1}}{2^{d-1} d v_d}, r_0\right)$-regular. Generally, we expect the c_0 satisfying Assumption 2 to diminish geometrically in d. However, as we see, its value does not actually impact the regret rate in T, only the leading constant, which is already exponential in d even if c_0 were d-independent.

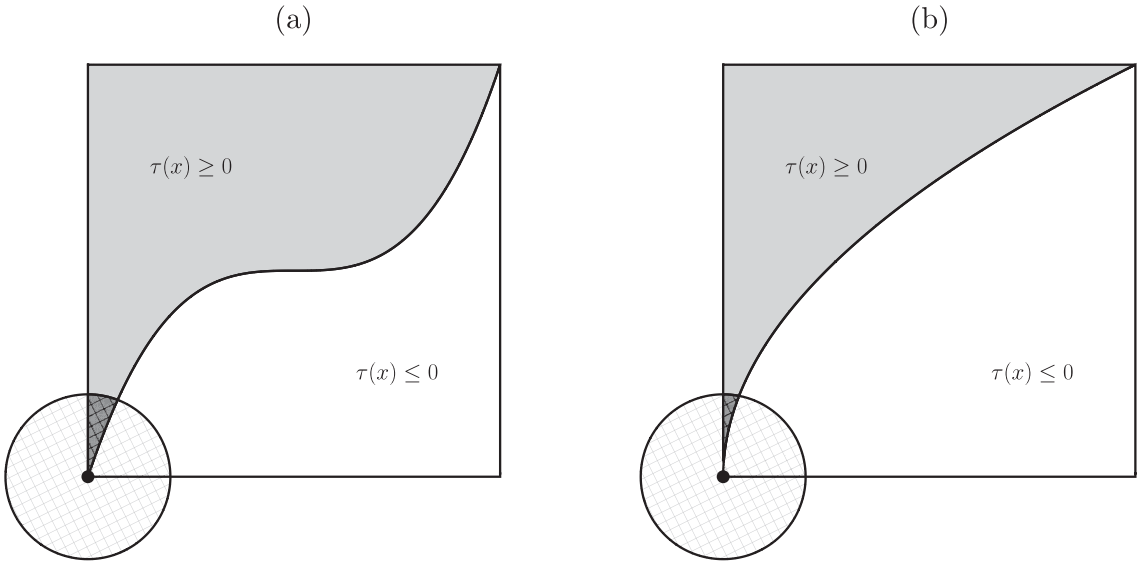
2.4. Bounded Covariate Density

Whereas Assumption 2 ensures there is sufficient volume around each point x at which we need to estimate $\eta_a(x)$, we also need to ensure that this translates to being able to collect sufficient data around each such point. Toward this end, we make the assumption that the contexts have a density and it is bounded away from zero and infinity.

Assumption 3 (Strong Density). *The marginal distribution of X has density $\mu(x)$ with respect to the Lebesgue measure, and μ is bounded away from zero and infinity on its support $\mathcal{X}$:*

$$0 < \mu_{\min} \leq \mu(x) \leq \mu_{\max} < \infty, \quad \forall x \in \mathcal{X}.$$

Figure 3. Illustration of Assumption 2



Notes. Each optimal decision region must be strongly regular in that the neighborhood of every point in the region must contain at least some constant fraction of the ball. (a) Assumption 2 is satisfied: every ball centered in $\mathcal{Q}_a$ has at least $c_0 = 1/12$ of its volume intersecting $\mathcal{Q}_a$. (b) Assumption 2 is violated: smaller balls centered at the corner have a vanishing fraction of their volume intersecting $\mathcal{Q}_a$.

Moreover, its support $\mathcal{X}$ is compact and $\mathcal{X} \subseteq [0,1]^d$.

Note that restricting $\mathcal{X}$ to $[0,1]^d$ is without loss of generality, having assumed compactness. Scaling and shifting the covariates to be in $[0,1]$ only affects the constants L, L_1 in Assumption 1.

Together, Assumptions 2 and 3 imply a lower bound on the probability that each arm is optimal.

Lemma 1. *Under Assumptions 2 and 3, we have $\mathbb{P}(X \in \mathcal{Q}_a) \geq p$ for $a = \pm 1$, where*

$$p = \mu_{\min} c_0 r_0^d v_d.$$

2.5. Margin Condition

We further impose a margin condition commonly used in stochastic contextual bandits (Rigollet and Zeevi 2010, Goldenshluger and Zeevi 2013) and classification (Mammen and Tsybakov 1999, Tsybakov 2004), which determines how the estimation error of expected rewards translates into regret of decision making.

Assumption 4 (Margin Condition). *The conditional average treatment effect function τ satisfies the margin condition with parameters $\alpha \geq 0$ and γ:*

$$\mathbb{P}(0 < |\tau(X)| \leq t) \leq \gamma t^\alpha \quad \forall t > 0.$$

The margin condition quantifies the concentration of contexts very near the decision boundary, at which the optimal action transitions from one arm to the other. This measures the difficulty of determining which of the two arms is optimal. When α is very small, the CATE function can be arbitrarily close to zero with high probability, so even very small estimation error of the CATE function may lead to suboptimal decisions. In contrast, when α is very large, the probability that expected arm rewards are very close to one another but not equal is very low, or, in other words, the expected rewards for two arms are nicely separated on most of $\mathcal{X}$.

2.6. Minimax Regret

Having now defined the problem and our assumptions about the distribution $\mathbb{P}$ defining the problem instance, we can introduce the notion of minimax regret. The minimax regret is the minimum over admissible policies π of the maximum of the regret of π over all problem instances $\mathbb{P}$ that fit our assumptions. This describes the best achievable behavior in the problem class we consider.

Formally, for $\beta \geq 1$, we let $\mathcal{P}(\beta, L_1, L, c_0, r_0, \mu_{\min}, \mu_{\max}, \gamma, \alpha)$ be the set of all distributions $\mathbb{P}$ on $(X, Y(-1), Y(+1)) \in [0,1]^d \times \mathbb{R} \times \mathbb{R}$ that satisfy Assumptions 1–4 with these parameters. For brevity, we write $\mathcal{P}$, implicitly considering the parameters as fixed. Letting Π denote all admissible policies, for some fixed parameters specifying a class $\mathcal{P}$, we then define the minimax regret as

$$\mathcal{R}_T = \inf_{\pi \in \Pi} \sup_{\mathbb{P} \in \mathcal{P}} R_T(\pi).$$

The minimax regret exactly characterizes how well we can hope to do in the given class of instances. It can be thought of as a game against nature in which nature plays second after we choose a policy, but we

know the set of plays available to nature (i.e., the instance class $\mathcal{P}$ with given parameters). Restricting the class is crucial for characterizing the dependence of regret on smoothness because the minimax regret against a single instance is always zero and the minimax regret against the class of instances with arbitrary β is linear in T. The minimax regret, therefore, characterizes the best achievable regret if one were only told the smoothness parameter (and additional preceding parameters) but the instance might be adversarially bad in every other way.

Now, we describe a general strategy for computing the minimax regret rate, which we follow in this paper. Suppose that, on the one hand, we can find a function $f(T)$ and an admissible policy $\hat{\pi}$ such that its regret for every instance $\mathbb{P} \in \mathcal{P}$ is bounded by the *same* function, $R_T(\hat{\pi}) \le f(T)$. Next, suppose that, on the other hand, we can show that there exists a function $f'(T) = \tilde{\Omega}(f(T))$, where, for every admissible policy π', there exists an instance $\mathbb{P} \in \mathcal{P}$ such that the regret is lower bounded by this same function, $R_T(\pi') \ge f'(T)$. Then we have shown two critical results: (a) the minimax regret satisfies the rate $\mathcal{R}_T = \tilde{\Theta}(f(T))$, and (b) we have a specific algorithm $\hat{\pi}$ that can actually achieve this best possible worst-case regret in rate, which also means the regret of $\hat{\pi}$ is known to be bounded in this rate for every single instance encountered.

In this paper, we proceed exactly as in the preceding. First, in Section 3, we develop a novel algorithm that can adapt to every smoothness level. Then, in Section 4.1 we prove a bound on its regret in every instance. Because this bound depends only on the parameters of $\mathcal{P}$, we, in fact, establish an upper bound on the minimax regret as before. In Section 4.2 we find a bad instance for every policy that yields a matching (up to polylogs) lower bound on its regret, establishing a lower bound on the minimax regret. This exactly yields the desired conclusion: a characterization of the minimax regret and the construction of a specific algorithm that achieves it.

2.7. On the Relationship of Margin and Smoothness

Before proceeding to develop a bandit algorithm for the smooth bandit problem and characterizing the minimax regret, we comment on the relationship between the smoothness of expected rewards and the margin assumption. Assumption 1 implies that the CATE function $\tau(x)$ is a member of the $(\beta, 2L, \mathcal{X})$-Hölder class with $\beta \ge 1$. Intuitively, when $\tau(x)$ is smooth, it cannot change too abruptly at the decision boundary $\tau(x) = 0$, so if it either touches or crosses the decision boundary at all, the mass near it must be significant (small α).

First, we present a direct corollary of proposition 3.4 of Audibert and Tsybakov (2005), who study (offline) classification with a smooth conditional probability function.

Proposition 1. *Suppose Assumptions 1–4 hold with $\alpha > 1$. Then, for all $x \in \text{interior}(\mathcal{X})$, there exists $r > 0, \sigma \in \{-1, 1\}$ such that $\sigma\tau(x') \ge 0$ for all $\|x' - x\| \le r$.*

Recall Assumptions 1–4 specify the class of the bandit problem we consider, so Proposition 1 is a statement about the instances in this class. Proposition 1 shows that, for a smooth bandit problem when $\alpha > 1$, all interior points have a neighborhood in which $\tau(x)$ is only nonnegative or only nonpositive, meaning $\tau(x)$ does not cross zero. Notice that, by continuity of τ, this also implies that, if any x and x' are in the same connected component of the interior (i.e., are connected by a interior path) then $\tau(x)\tau(x') \ge 0$ so that there must exist an optimal policy $\pi^*(x)$ in Equation (1) that is constant on connected components of the interior. However, $\tau(x)$ might still be arbitrarily close to zero, especially as we vary the instance in the class of instances $\mathcal{P}$ to compute the minimax regret, potentially making it difficult to distinguish the optimal arm and still requiring nontrivial regret.

We next show that this, however, does not happen when the margin is very large.

Proposition 2. *Suppose Assumptions 1–4 hold with $\alpha > d$. Then, there exists a positive constant $\tau_{\min}$ depending only on the parameters of Assumptions 1–4 such that, for any $x \in \mathcal{X}$, we have either $\tau(x) = 0$ or $|\tau(x)| \ge \tau_{\min}$.*

By continuity of $\tau(x)$, Proposition 2 implies that, on each connected component of $\mathcal{X}$, $\tau(x)$ has a constant sign (negative, zero, or positive). In particular, as it would contradict Lemma 1, Proposition 2 implies that there exist no smooth bandit instances with $\alpha > d$, $\mathcal{X}$ connected and $\tau(x)$ not the constant zero function on $\mathcal{X}$ as such would require $\mathbb{P}(X \in \mathcal{Q}_1) \wedge \mathbb{P}(X \in \mathcal{Q}_{-1}) = 0$.

More crucially, Proposition 2 makes an implication on minimax regret when $\alpha > d$ because $\tau_{\min}$ is a *uniform* bound (and, in this sense, the result is stronger than the statement corresponding to $\alpha > d$ in proposition 3.4 of Audibert and Tsybakov 2005). Notice that $|\tau(X)| \in \{0\} \cup [\tau_{\min}, \infty)$ implies that Assumption 4 holds for any $\alpha \ge 0$ (simply let $\gamma = \tau_{\min}^{-\alpha}$). Recall that the class of instances $\mathcal{P}$ in Section 2.6 is defined in terms of the parameters of Assumptions 1–4. Therefore, Proposition 2 shows that, for any $\alpha' \ge \alpha > d$ (and $\gamma \ge \tau_{\min}^{-\alpha'}$ sufficiently large), the minimax regret in the class of instances $\mathcal{P}$ is upper bounded by the minimax regret in the class $\mathcal{P}'$, where we set α to the *larger* α'. More to the point, in the following, by exhibiting a feasible algorithm, we establish an upper bound on minimax regret of $\tilde{O}(1)$ whenever $\alpha \ge 1 + d/\beta$. Proposition 2 shows that the same $\tilde{O}(1)$ bound applies even if just $\alpha > d \wedge (1 + d/\beta)$.

3. SmoothBandit: A Low-Regret Algorithm for Any Smoothness Level

In this section, we develop our algorithm, SmoothBandit (Algorithm 1). Various design choices in our algorithm are made with an eye toward tractably analyzing its regret. In Section 5, we propose a simplified version, albeit heuristic and lacking analysis, and use it to conduct numeric experiments that demonstrate our theory correctly predicts regret behavior in practice.

We first review local polynomial regression, which we use in our algorithm to estimate η_a.

3.1. The Local Polynomial Regression Estimator

A standard result of (off-line) nonparametric regression is that the smoother a function is in terms of its Hölder parameter β, the faster it can be estimated. Appropriate convergence rates can, for example, be achieved using local polynomial regression estimators that adjust to different smoothness levels (Stone 1980, 1982). In this section, we briefly review local polynomial regression and its statistical property in an *off-line* bandit setting. Its use in our *online* algorithm is described in Section 3.2. More details about local polynomial regression can be found in Tsybakov (2008) and Audibert and Tsybakov (2007).

Consider an off-line setting, in which we have access to an exogenously collected i.i.d. sample, $S=\{(X_t,Y_t)\}_{t=1}^n$ drawn i.i.d. from (X, Y), where X has support $\mathcal{X}\subset\mathbb{R}^d$. We can then estimate the regression function $\eta(x)=\mathbb{E}[Y|X=x]$ at every point x using the following local polynomial estimator.

Definition 3 (Local Polynomial Regression Estimator). For any $x\in\mathcal{X}$, given a bandwidth $h>0$, an integer $l\geq 0$, samples $S=\{(X_t,Y_t)\}_{t=1}^n$, and a degree-l polynomial model $\theta(u;x,\vartheta,l)=\sum_{|r|\leq l}\vartheta_r(S)(u-x)^r$, define the local polynomial estimate for $\eta(x)$ as $\hat{\eta}^{\mathrm{LP}}(x;S,h,l)=\theta(x;x,\hat{\vartheta}_x,l)$, where

$$\hat{\vartheta}_x\in\arg\min_{\vartheta}\sum_{t:X_t\in\mathcal{B}(x,h)}(Y_t-\theta(X_t;x,\vartheta,l))^2. \tag{4}$$

For concreteness, we define $\hat{\eta}^{\mathrm{LP}}(x;S,h,l)=0$ if the minimizer is not unique.

In words, the local polynomial regression estimator fits a polynomial by least squares to the data that is in the h-neighborhood of the query point x and evaluates this fit at x to predict $\eta(x)$.

Because Equation (4) is a least squares problem, the estimation accuracy of the local polynomial estimator $\hat{\eta}^{\mathrm{LP}}(x;S,h,l)$ depends on the associated Gram matrix:

$$\hat{\mathcal{A}}(x;S,h,l)=\{\hat{\mathcal{A}}_{r_1,r_2}(x;S,h)\}_{|r_1|,|r_2|\leq l},$$

$$\text{where}\quad \hat{\mathcal{A}}_{r_1,r_2}(x;S,h)=\sum_{t:X_t\in\mathcal{B}(x,h)}\left(\frac{X_t-x}{h}\right)^{r_1+r_2}. \tag{5}$$

The following proposition illustrates (using the off-line setting as an example) why our Assumptions 1–3 are crucial in our problem. In particular, it shows that bounded density and strong regularity of the support of the data ensure a well-conditioned, locally weighted Gram matrix. Moreover, it shows how the bandwidth and polynomial degree should adapt to the smoothness level β. This proposition is a direct extension of theorem 3.2 of Audibert and Tsybakov (2007). We include this result purely for motivation, whereas in our online setting, we need to establish a more refined result that accounts for our adaptive data collection.

Proposition 3. *Let S be an i.i.d. sample of (X, Y), where η is $(\beta,L,\mathcal{X})$-Hölder, $\mathcal{X}$ is compact and strongly (c_0, r_0)-regular, and X has a density bounded away from zero and infinity on $\mathcal{X}$. Then, there exist positive constants λ_0,C_1,C_2 such that, for any $x\in\mathcal{X}$ and $\epsilon>0$ with probability at least $1-C_1\exp\left\{-C_2 n_a^{\frac{2\beta}{2\beta+d}}\varepsilon^2\right\}$, we have*

$$\lambda_{\min}(\hat{\mathcal{A}}(x;S,n^{-1/(2\beta+d)},\mathfrak{b}(\beta)))\geq\lambda_0,$$

$$\text{and}\quad |\hat{\eta}^{\mathrm{LP}}(x;S,n^{-1/(2\beta+d)},\mathfrak{b}(\beta))-\eta(x)|\leq\varepsilon.$$

In our online bandit setting, the samples for each arm are collected in an adaptive way because both exploration and exploitation can depend on data already collected. As a result, the distribution of the samples for each arm is considerably more complicated. Thus, we need to use the local polynomial estimator in a somewhat more sophisticated way and analyze it more carefully. See Sections 3.2 and 4.1 for the details.

3.2. Our Algorithm

Algorithm 1 (SmoothBandit)

Input: Grid lattice G, epoch schedule $\{\mathcal{T}_k\}_{k=1}^K$, Hölder smoothness constant β, strong regularity constant c_0, context dimension d, context support $\mathcal{X}$.

1: Initialize $\mathcal{E}_{\pm1,1}=\emptyset,\mathcal{R}_1=\mathcal{X}$ (exploit nowhere, explore everywhere)
2: **for** $t\in\mathcal{T}_1$, **do**
3: Pull $A_t=\pm1$ randomly, equiprobably
4: end for
5: Log the samples $S_{\pm1,1}=\{(X_t,Y_t):t\in\mathcal{T}_1,A_t=\pm1\}$
6: **for** $k=2,3,\ldots,K$, **do**
7: Identify inestimable regions for local polynomial regression with bandwidth $H_{a,k-1}$ $(a=\pm1)$:

$$\mathcal{D}_{a,k-1}=\bigcup\left\{\mathrm{Cube}(x):x\in\mathcal{R}_{k-1}\cap G,\ \left(\bigcup_{j=1}^{k-1}\mathcal{E}_{a,j}\cup\mathcal{R}_{k-1}\right)\cap\mathcal{X}\text{ is not weakly }\left(\frac{c_0}{2^d},H_{a,k-1}\right)\text{-regular at }x\right\} \tag{6}$$

8: Set $N_{\pm1,k-1}=|S_{\pm1,k-1}|,H_{\pm1,k-1}=N_{\pm1,k-1}^{-1/(2\beta+d)}$

9: Construct the CATE estimate for every $x \in G \cap \mathcal{R}_{k-1} \cap \mathcal{D}^C_{1,k-1} \cap \mathcal{D}^C_{-1,k-1}$,

$$\hat{\tau}_{k-1}(x) = \hat{\eta}^{\mathrm{LP}}(x; S_{+1,k-1}, H_{+1,k-1}, \mathfrak{b}(\beta)) - \hat{\eta}^{\mathrm{LP}}(x; S_{-1,k-1}, H_{-1,k-1}, \mathfrak{b}(\beta)) \quad (7)$$

10: Update decision regions: for $a = \pm 1$,

$$\mathcal{E}_{a,k} = \bigcup \{\mathrm{Cube}(x) : x \in G \cap \mathcal{R}_{k-1} \cap \mathcal{D}^C_{1,k-1} \cap \mathcal{D}^C_{-1,k-1}, a\hat{\tau}_{k-1}(x) > \epsilon_{k-1}\} \cup \mathcal{D}_{-a,k-1}, \quad (8)$$

$$\mathcal{R}_k = \bigcup \{\mathrm{Cube}(x) : x \in G \cap \mathcal{R}_{k-1} \cap \mathcal{D}^C_{1,k-1} \cap \mathcal{D}^C_{-1,k-1}, |\hat{\tau}_{k-1}(x)| \le \epsilon_{k-1}\}. \quad (9)$$

11: **for** $t \in \mathcal{T}_k$, **do**
12: **if** $X_t \in \bigcup_{j=1}^k \mathcal{E}_{+1,j}$, **then** pull $A_t = +1$
13: **else if** $X_t \in \bigcup_{j=1}^k \mathcal{E}_{-1,j}$, **then** pull $A_t = -1$
14: **else** pull $A_t = \pm 1$ randomly, equiprobably
15: **end for**
16: Log samples $S_{\pm 1,k} = \{(X_t, Y_t) : t \in \mathcal{T}_k, A_t = \pm 1\}$
17: **end for**

In this section, we present our new algorithm for smooth contextual bandits, which uses local polynomial regression estimators that adjust to any smoothness level. The algorithm is summarized in Algorithm 1. We review its salient features. In what follows, we assume a fixed horizon T but can accommodate an unknown, variable T using the well-known doubling trick (see Auer et al. 1995, Cesa-Bianchi and Lugosi 2006).

3.2.1. Algorithm Overview. We begin with a rough sketch of the overall structure of the algorithm. Specifics are given in Algorithm 1 and in the sections that follow. Our algorithm makes a cover $\mathcal{C}$ of the covariate support using a grid of hypercubes, $\mathcal{X} = \bigcup_{S \in \mathcal{C}} S$, where $\mathcal{C}$ consists of the intersections of $\mathcal{X}$ with disjoint half-open, half-closed hypercubes with a finely tuned side length (see Section 3.2.2). Our algorithm then proceeds in epochs of (roughly) geometrically increasing time lengths (see Section 3.2.3). At the beginning of the kth epoch, each cube $S \in \mathcal{C}$ in the cover is assigned either to the random exploration region, $\mathcal{R}_k$, or to one of two *exploitation* regions, $\mathcal{E}_{+1,k}, \mathcal{E}_{-1,k}$ (see Section 3.2.6). During the kth epoch, if X_t falls in $\mathcal{R}_k$, we pull arms ± 1 with probability 1/2 each, and if X_t falls in $\mathcal{E}_{a,k}$, we pull arm a. We start out with randomizing everywhere, $\mathcal{R}_0 = \mathcal{X}$. Then, as we collect more data, we peel hypercubes away from the randomization region and into the exploration region, in which we declare one of the two arms is almost certainly optimal based on observations from the epoch that just concluded. There are two ways to infer that a hypercube should be moved to an exploitation region: either we have already declared one of the arms is almost certainly optimal in very many nearby hypercubes (see Section 3.2.5), or we have enough data near the center of the hypercube x_0 to fit a high-fidelity local polynomial regression estimate for both $\eta_{+1}(x_0), \eta_{-1}(x_0)$ (this involves data *outside* the hypercube), and the difference is large enough to rule out (with high probability) that one arm appears better only because of estimation noise, so we declare the apparently dominant arm is indeed dominant (see Section 3.2.4). Thus, we maintain a plan of action of how we will act in each round t depending on the observed context X_t, and at the end of each epoch, we update this plan by declaring more context regions as exploitation regions in which we only pull one of the two arms. This structure is mimicked by our multi-arm extension in which we maintain an active set of arms (subset of $\mathcal{A}$) for each hypercube (see Online Appendix A).

3.2.2. Grid Structure. Following Stone (1982) and similarly to previous nonparametric-response bandit literature (Rigollet and Zeevi 2010, Perchet and Rigollet 2013), we partition the context space into small hypercubes. For each time step, both our estimates of $\eta_a(x)$ and our policy $\pi_t(x)$ are piecewise constant on these hypercubes. Specifically, in each hypercube, we either pull arm +1, pull arm –1, or equiprobably pull a random arm from among the two (see Figure 4(c)). Crucially, and differently from Rigollet and Zeevi (2010) and Perchet and Rigollet (2013), we use data from both inside and outside each hypercube to define the estimates and action inside each hypercube. In particular, these hypercubes are *different* from the *bandwidth* that we use for estimation, which is orders of magnitude larger.

We first define a grid lattice G' on $[0,1]^d$: letting $\delta = T^{-\frac{\beta}{2\beta+d}} (\log T)^{-1}$,

$$G' = \left\{ \left(\frac{2j_1+1}{2}\delta, \ldots, \frac{2j_d+1}{2}\delta \right) : j_i \in \{0, \ldots, \lceil \delta^{-1} \rceil - 1\}, i = 1, \ldots, d \right\}.$$

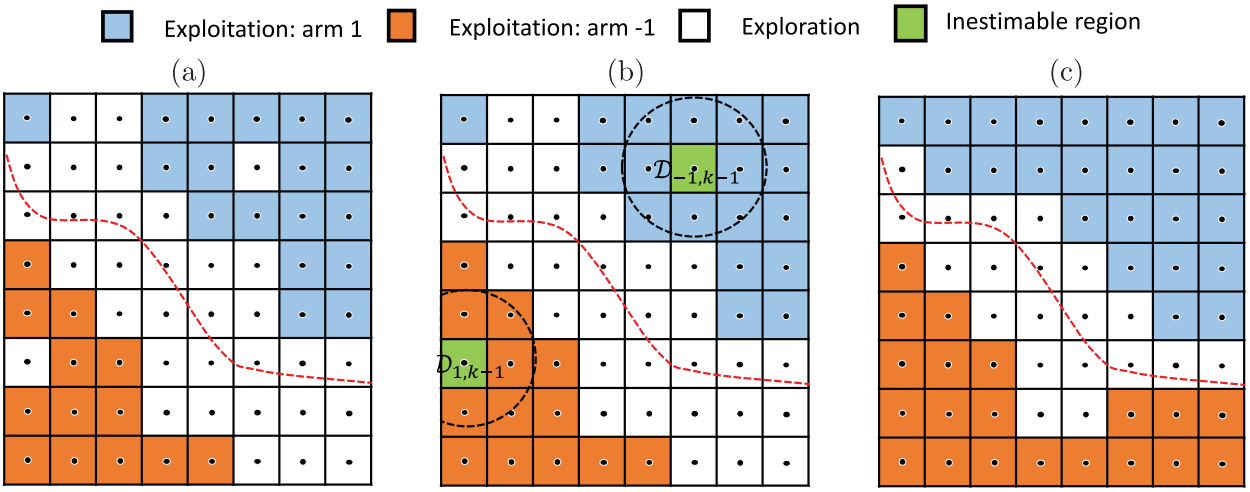
For any $x \in \mathcal{X} \subseteq [0,1]^d$, we denote by $g(x) = \arg\min_{x' \in G'} \|x - x'\|$ the closest point to x in G'. If there are multiple closest points to x, we choose $g(x)$ to be the one closest to $(0, 0, \cdots, 0)$. All points that share the same closest grid point $g(x)$ belong to a hypercube with length δ and center $g(x)$. We denote this hypercube as $\mathrm{Cube}(x) = \{x' \in \mathcal{X} : g(x') = g(x)\}$ and the collection of all such hypercubes overlapping with the covariate support as

$$\mathcal{C} = \{\mathrm{Cube}(x) : x \in G\}, \quad \text{where}$$

$$G = \{x \in G' : \mathbb{P}(\mathrm{Cube}(x) \cap X) > 0\}.$$

Note that the union of all cubes in $\mathcal{C}$, $\bigcup_{S \in \mathcal{C}} S$ must cover the covariate support $\mathcal{X} \subseteq [0,1]^d$.

Figure 4. (Color online) Updating the Decision Regions



Notes. The black dots represent hypercube centers given in G. The dashed curve is the true decision boundary. (a) The decision regions at the end of the $k-1^{\text{th}}$ epoch: each hypercube is assigned with one action, either always pull arm 1, always pull arm –1, or pull one of the arms at random equiprobably. (b) In epoch k, we cannot sufficiently accurately estimate η_a in $\mathcal{D}_{a,k-1}$ for lack of sufficient samples from arm a in the local neighborhoods (black dashed circle). (c) The decision regions at the end of k^{th} epoch: previous exploration regions are moved to exploitation either because of large estimated CATE when estimable or because of inestimability.

3.2.3. Epoch Structure. Our algorithm then proceeds in an epoch structure, in which the estimates and actions assigned to each hypercube are fixed for the duration of that epoch. For each epoch, we target a CATE-estimation error tolerance of $\epsilon_k = 2^{-k}$. With this target in mind, we set the length of the k^{th} epoch as follows:

$$n_k = \left\lceil \frac{4}{p}\left(\frac{\log(T\delta^{-d})}{C_0\epsilon_k^2}\right)^{\frac{2\beta+d}{2\beta}} + \frac{2}{p^2}\log T \right\rceil, \tag{10}$$

where p, C_0 are positive constants given in Lemmas 1 and 7, respectively. We further denote the time index set associated with the k^{th} epoch as $\mathcal{T}_k = \{t : \sum_{i=1}^{k-1} n_{i-1} + 1 \le t \le \min\{\sum_{i=1}^{k} n_i, T\}\}$.

In our algorithm, we continually maintain a growing region, comprising hypercubes, in which we are near-certain which of the arms is optimal. In these regions, we always pull the seemingly optimal arm. In contrast, we randomize whenever we are not sure (denoted by the region $\mathcal{R}_k$ for epoch k). The first epoch, $\mathcal{T}_1$, is a cold-start phase in which, lacking any information, we simply pull each arm uniformly at random in every hypercube ($\mathcal{R}_1 = \mathcal{X}$). After that point, once we have some data, for each subsequent epoch, $k \ge 2$, we add the hypercubes $\mathcal{E}_{a,k} \subseteq \mathcal{R}_{k-1}$ to the set of hypercubes in which we just learned that arm a is probably optimal, never removing any hypercube that was before added. This means that, in epoch k, we are collecting data on arm a exclusively in the region $\bigcup_{j=1}^{k} \mathcal{E}_{a,j} \cup \mathcal{R}_k$. We describe in detail how we determine which hypercubes, $\mathcal{E}_{a,k}$, to add to the exploitation region of each arm in each epoch in Sections 3.2.5 and 3.2.6.

The total number of epochs K is the minimum integer such that $\sum_{k=1}^{K} n_k \ge T$. The following lemma shows that K grows at most logarithimically with T under the epoch schedule in Equation (10).

Lemma 2. *When* $T \ge e^{C_0 \vee 1}$,

$$K \le \left\lceil \frac{\beta}{(2\beta+d)\log 2}\log(T) \right\rceil.$$

3.2.4. Estimating CATE. Next, we describe how we estimate the expected rewards, $\eta_{\pm 1}(x)$, and CATE, $\tau(x) = \eta_1(x) - \eta_{-1}(x)$, which we use to determine the action we take in each hypercube in each epoch. In particular, at the start of each k^{th} epoch, $k \ge 2$, we estimate each arm's expected reward $\eta_a(x)$ using the data for each arm from the last epoch, which we denote by $S_{a,k-1}$ as in Algorithm 1. Our proposed estimate is the following piecewise constant modification of the local polynomial regression estimate:

$$\hat{\eta}_{a,k-1}(x) = \hat{\eta}^{\text{LP}}(g(x); S_{a,k-1}, H_{a,k-1}, \mathfrak{b}(\beta)), \quad \text{where} \tag{11}$$
$$H_{a,k-1} = N_{a,k-1}^{-1/(2\beta+d)}, \quad N_{a,k-1} = |S_{a,k-1}|.$$

Note that, by construction, $\hat{\eta}_{a,k-1}(x) = \hat{\eta}_{a,k-1}(x')$ whenever x and x' belong to the same hypercube $g(x) = g(x')$. Then, our CATE estimate, $\hat{\tau}_{k-1}(x)$, is simply the difference of these for $a = \pm 1$. Because we only evaluate $\hat{\tau}_{k-1}(\cdot)$ at hypercube centers $x \in G$ in our Algorithm 1,

we simply use $g(x)$ as the argument to the local polynomial estimates in Equation (7). In particular, we only need to compute two local polynomial regression estimates at a subset of the (finitely many) grid points. Note that some grid points may not even belong to $\mathcal{X}$ because their hypercubes may not be fully contained in $\mathcal{X}$; nevertheless, we can use these centers as representative as their $H_{\pm 1,k-1}$ neighborhood still contains sufficient data (Lemma 3). Note also that the associated sample sizes, $N_{\pm 1,k-1}$, are random variables because they depend on how many samples in the $(k-1)^{\text{th}}$ epoch fall in different decision regions and on the random decision regions themselves.

Similar to the nondifferentiable bandit of Rigollet and Zeevi (2010), our estimators, Equation (11), are hypercube-wise constant. That the estimate at the center of each hypercube is a good estimate for the whole hypercube is justified by the smoothness of $\eta_{\pm 1}$, and the error is controlled by the size of the hypercubes (see Lemma EC.10 for details).

However, differently from Rigollet and Zeevi (2010), our estimate at the center of each hypercube uses data from both inside and outside the hypercube instead of only inside. This is established by the next lemma. (Recall that our Assumption 1 requires $\beta \geq 1$.)

Lemma 3. *There exists a positive constant c_1 such that*

$$\frac{H_{\pm 1,k}}{\sqrt{d}\delta} \geq c_1 T^{\beta-1} \log(T)^{\frac{(2\beta-1)(2\beta+d)}{2\beta}}.$$

When $\beta \geq 1$, there exists $T_0 > 0$ such that, for $T > T_0$, $H_{\pm 1,k} \geq \sqrt{d}\delta$ for $1 \leq k \leq K$.

Lemma 3 shows that the bandwidth we use, that is, the neighborhood of data used to construct the estimate, is *much larger* than the hypercube size, with which the estimate is used. Note that although the variable $H_{\pm 1,k}$ is random, the statement in Lemma 3 is always true. According to the nonparametric estimation literature (Stone 1980, 1982), the proposed hypercube size and bandwidths (up to logarithmic factors) are crucial for achieving optimal nonparametric estimation accuracy for smooth functions. This means we indeed *need* to leverage the more global information in order to leverage the smoothness of expected reward functions. This also means that separating the problems into isolated MABs within each hypercube, as would be optimal for unsmooth rewards, is infeasible: we must use data across hypercubes to be efficient, and so decisions in different hypercubes are interdependent. In particular, our actions in one hypercube affect how many samples we collect to learn rewards in other hypercubes.

3.2.5. Screening Out Inestimable Regions and Accuracy Guarantees.

Although using data across multiple hypercubes enables us to improve the estimation accuracy for smooth expected reward functions, it also introduces complicated dependence between data collection in *one* hypercube and algorithm decisions in *other* hypercubes. More concretely, the number of samples available to estimate η_a in each hypercube and, correspondingly, the accuracy of this estimate depends on the arms we pull in other, neighboring hypercubes. Because, in each epoch in each hypercube, we either always exploit or randomly explore, this problem arises precisely when there is a hypercube that is surrounded by hypercubes in which we are sure about the optimal arm (and, therefore, do not explore both arms) but in which we are not yet sure about the optimal arm (and, therefore, need to estimate both arm reward functions). (See Figure 4, (b) and (c).) As a result, the local polynomial regression for estimating η_a in this hypercube can be ill-conditioned and fail to ensure our accuracy target ϵ_k. Worse yet, this problem continues to persist at all future epochs because the nearby hypercubes continue to exploit, and the accuracy target only becomes sharper.

Luckily, it turns out that, whenever such a problem arises, we do not actually need to estimate η_a in these hypercubes: the fact that the hypercube is surrounded by neighboring hypercubes in which we are sure one arm is optimal means that the same arm is also optimal in this hypercube with high probability (see Lemma 5). The only thing we need to do is to detect this issue correctly. Specifically, we propose to use the rule in Equation (6) in order to screen out the inestimable regions. This screening rule is motivated by Proposition 3 and Assumption 2, which imply that the strong regularity property of the support of the samples $S_{a,k-1}$ (i.e., $(\bigcup_{j=1}^{k-1} \mathcal{E}_{a,j} \cup \mathcal{R}_{k-1}) \cap \mathcal{X}$) is critical for the conditioning of the local polynomial estimator. We show in Lemma EC.9 that this screening procedure is well-defined: any hypercube in $\mathcal{C}$ can be classified into at most one of $\mathcal{D}_{1,k}$ and $\mathcal{D}_{-1,k}$ but not both. Moreover, although we check only *weak* $(\frac{c}{2^d}, H_{a,k-1})$-regularity with respect to only hypercube centers, Lemma EC.8 implies a far stronger consequence for the proposed screening rule: $(\bigcup_{j=1}^{k-1} \mathcal{E}_{a,j} \cup \mathcal{R}_{k-1}) \cap \mathcal{X}$ is not *strongly* (c_0, r_0)-regular at any point in $\mathcal{D}_{a,k-1}$.

After removing these inestimable regions, we can show (Theorem 1) that our uniform estimation error anywhere in the remaining uncertain region from each epoch (i.e., $\mathcal{R}_k \cap D_{1,k}^C \cap D_{-1,k}^C$) is *exponentially* shrinking:

$$\sup_{x \in \mathcal{R}_k \cap D_{1,k}^C \cap D_{-1,k}^C} |\hat{\tau}_k(x) - \tau(x)| \leq \epsilon_k \text{ with probability } 1 - O(T^{-1}). \tag{12}$$

3.2.6. Decision Regions.

We start by randomizing everywhere, $\mathcal{R}_1 = \mathcal{X}$, and in each subsequent epoch, we

remove the hypercubes $\mathcal{E}_{-1,k}, \mathcal{E}_{1,k}$ from the randomization region $\mathcal{R}_k$ and assign them to join the growing exploitation regions. The set $\mathcal{E}_{a,k}$ is the union of two parts. The first, $\{x \in \mathcal{R}_{k-1} \cap \mathcal{D}^C_{1,k-1} \cap \mathcal{D}^C_{-1,k-1} : a\hat{\tau}_{k-1}(x) > \epsilon_{k-1}\}$, is determined by $\hat{\tau}_{k-1}$ and consists of the points at which, as long as the event in Equation (12) holds, we are sure arm a is optimal. The second is $\mathcal{D}_{-a,k-1}$ and, in contrast to the first, we cannot rely on the CATE estimator in order to determine that a is optimal here. Nevertheless, we can show that $\mathcal{D}_{-a,k-1} \cap \mathcal{X} \subseteq \{x \in \mathcal{X} : a\tau(x) > 0\}$ under Assumption 2 and as long as the event in Equation (12) holds (Lemma 5). This means that we can conclude that the arm a is also optimal on $\mathcal{D}_{-a,k-1}$ even though we cannot estimate CATE accurately there.

The remaining randomization region in each epoch, $\mathcal{R}_k$, consists of the subset of the previous randomization region in which we cannot determine that either arm is optimal using either of these criteria. In particular, the CATE estimate is below the accuracy target inside $\mathcal{R}_k$, $|\hat{\tau}_{k-1}(x)| \le \epsilon_{k-1}$, so even when the event in Equation (12) holds, we cannot be sure which arm is optimal. Thus, we may as well pull each arm uniformly at random to provide maximum exploration for estimation in future epochs. Moreover, the exploration cost is manageable because, as long as the event in Equation (12) holds: (1) the regret incurred from pulling suboptimal arms at the randomization region shrinks exponentially because $|\tau(x)| \le |\hat{\tau}_{k-1}(x)| + \epsilon_{k-1} \le 2\epsilon_{k-1}$ for $x \in \mathcal{R}_k$, and (2) the randomization region shrinks over the epochs as Assumption 4 implies that $\mathbb{P}(X \in \mathcal{R}_k \cap \mathcal{X}, |\tau(X)| \neq 0) \le \mu(\{X : 0 < |\tau(X)| \le 2\epsilon_{k-1}\}) \le \gamma(2\epsilon_{k-1})^\alpha$.

In each epoch, we update the CATE estimates and the decision rule only when it is needed. We estimate CATE and design new decision regions (i.e., $\mathcal{R}_k$ and $\mathcal{E}_{\pm 1,k}$) only within the region in which we failed to learn the optimal arm with high confidence in previous epochs (i.e., $\mathcal{R}_{k-1}$), and we follow previous decision rules on regions in which the optimal arm is already learned with high confidence (i.e., $\bigcup_{j=1}^{k-1} \mathcal{E}_{a,j}$). In this way, we gradually refine the accuracy of the CATE estimator in ambiguous regions while making efficient use of the information learned in previous epochs.

3.2.7. Comparison with (A)BSE When $\beta = 1$. Whereas our algorithm is most notable for tackling the case of $\beta > 1$, it also handles the special case of $\beta = 1$, which is exactly the intersection point with the previous literature that focuses on $\beta \le 1$ (Rigollet and Zeevi 2010, Perchet and Rigollet 2013). Even when $\beta = 1$, our algorithm is distinct from these. Notably, UCBograms in Rigollet and Zeevi (2010) and binned successive elimination (BSE) in Perchet and Rigollet (2013), which both run isolated MAB algorithms in each hypercube, can achieve minimax optimal regret only when $\alpha \le 1$ (and $\beta \le 1$). Because these algorithms learn expected rewards using data only within each hypercube, they require pulling each arm at least once in each hypercube and, thus, necessarily incur regret of $\Omega(T^{\frac{d}{2+d}})$ when $\beta = 1$, which is suboptimal when $\alpha > 1$ because the minimax regret rate in this case is $\Theta(T^{\frac{1+\alpha+d}{2+d}})$. Indeed, when $\alpha > 1$, the expected rewards for two arms are relatively separated, and we can tell apart the optimal arms with relatively little data, so pulling each arm at least once in every hypercube may be wasteful. Addressing this is the central purpose of the ABSE algorithm in Perchet and Rigollet (2013), which gradually refines the hypercubes (though, still, it can only handle $\beta \le 1$).

Our algorithm provides another way around this issue by using data across hypercubes even in the special case of $\beta = 1$, in which our bandwidth is larger than the hypercube size in all but the last epoch. Then, whenever a particular hypercube has arms that are well-separated (as many must be when α is large), we can still detect this even if we did not pull both arms in this hypercube. For example, in the k^{th} epoch, Lemma 3 ensures that the learning radius $H_{\pm 1,k}$ is much larger than the hypercube size δ, so even if we have not pulled one of the arms in some hypercubes yet, we can still collect enough samples (with high probability) for both arms in their neighborhood within the learning radius so that we can construct the CATE estimator $\hat{\tau}_{k-1}$ that achieves the target precision level ϵ_{k-1} (see Theorem 1 for the formal statement). If the expected rewards for the two arms on some of these hypercubes are separated enough so that $|\hat{\tau}_{k-1}(x)| > \epsilon_{k-1}$ on them, then we can confidently push them into exploitation regions. As a result, we do not "waste" arm pulls in these hypercubes. Importantly, our algorithm can determine optimal arms on hypercubes with well-separated expected rewards in early epochs using relatively imprecise CATE estimators based on small samples and do so on more difficult hypercubes with less separated expected rewards later on using more precise CATE estimators. In this way, it carefully achieves the minimax optimal regret rate even when $\alpha > 1$ (see Theorem 2). For $\beta = 1$, the behavior of our algorithm is similar but different from ABSE in that both gradually refine the learning radius, but in ABSE, the learning radius is set to be the same as the hypercube size, whereas in our algorithm, the learning radius is different from the hypercube size.

3.2.8. Finite Running Time. Finally, we remark that Algorithm 1 can be run in finite time. First, we show that all decision regions are unions of hypercubes in $\mathcal{C}$ as shown in Figure 4.

Lemma 4. *For $1 \le k \le K$, $\mathcal{E}_{\pm 1,k}$, $\mathcal{D}_{\pm 1,k}$ and $\mathcal{R}_k$ are all unions of hypercubes in $\mathcal{C}$.*

The number of hypercubes itself, $|G|$, is of course finite. To determine in what hypercube an arriving context falls, we need only divide each of its coordinates by δ.

The remaining question is to compute which hypercubes belong to which decision region at the start of each epoch. To compute $\mathcal{D}_{\pm 1,k}$, we need to compute the volume in the intersection of $\mathcal{X}$, a union of cubes, and a ball and compare it to a given constant. We need to do this at most once in each epoch for every hypercube. If $\mathcal{X}$ has a simple shape, such as the unit hypercube, this can be done analytically. Alternatively, given a membership oracle for $\mathcal{X}$, we can compute this using rectangle quadrature integration. In particular, we can easily allow for some slowly vanishing approximation error in the quadrature integration without deteriorating the regret rate of our algorithm. Then, to compute $\mathcal{E}_{\pm 1,k}$ and $\mathcal{R}_k$, we need only to compute $\hat{\eta}_{k,a}(x)$ at most once in each epoch at each lattice point $x \in G$. Computing this estimate requires constructing an $M_\beta \times M_\beta$ matrix given by averaging over the data within the bandwidth neighborhood and then pseudo-inverting this matrix.

4. Theoretical Guarantees: Upper and Lower Bounds on Minimax Regret

We next provide two results that together characterize the minimax regret rate (up to polylogs): an upper bound on the regret of our algorithm, and a matching lower bound on the regret of any other algorithm (in the regime in which $\alpha\beta \le d$).

4.1. Regret Upper Bound

In this section, we derive an upper bound on the regret of our algorithm. The performance of our algorithm, as we show in this section, crucially depends on two events: $\mathcal{M}_k$, the event that sufficiently *many* samples for each arm are available for CATE estimation at the end of epoch k, and $\mathcal{G}_k$, the event that our estimator $\hat{\tau}_k$ has *good* accuracy.

Concretely,

$$\mathcal{M}_k = \left\{ N_{1,k} \wedge N_{-1,k} \ge \left(\frac{\log(T\delta^{-d})}{C_0 \epsilon_k^2} \right)^{\frac{2\beta+d}{2\beta}} \right\},$$

$$\mathcal{G}_k = \left\{ \sup_{x \in \mathcal{R}_k \cap D_{1,k}^C \cap D_{-1,k}^C} |\hat{\tau}_k(x) - \tau(x)| \le \epsilon_k \right\}.$$

For convenience, we also define $\overline{\mathcal{G}}_k = \cap_{1 \le j \le k} \mathcal{G}_j$ and $\overline{\mathcal{M}}_k = \cap_{1 \le j \le k} \mathcal{M}_j$, where an empty intersection ($\overline{\mathcal{G}}_0$ or $\overline{\mathcal{M}}_0$) is the whole event space (always true).

4.1.1. Characterization of the Decision Regions.

The following lemma shows that these two events are critical for the effectiveness of the proposed decision rules in that, whenever they hold, we have the desired behavior described in Sections 3.2.6 and 3.2.5.

Lemma 5. *Fix any $k \ge 1$. Suppose Assumption 2 holds and that $T \ge T_0 \vee \left(\exp\left(1 \vee \frac{C_0(2\beta+d)}{4(2r_0)^{2\beta}(2\beta+d+\beta d)} \right) \right)$ with T_0 given in Lemma 3 and C_0 given in Lemma 7. Then, under the event $\overline{\mathcal{G}}_{k-1} \cap \overline{\mathcal{M}}_{k-1}$, we have for $a = \pm 1$:*

i. $\mathcal{R}_k \cap \mathcal{X} \subseteq \{x \in \mathcal{X} : |\tau(x)| \le 2\epsilon_{k-1}\}$.
ii. $(\cup_{j=1}^k \mathcal{E}_{a,j}) \cap \mathcal{X} \subseteq \{x \in \mathcal{X} : a\tau(x) > 0\}$.
iii. $\mathcal{Q}_a \subseteq ((\cup_{j=1}^k \mathcal{E}_{a,j}) \cup \mathcal{R}_k) \cap \mathcal{X}$.
iv. $\mathcal{D}_{a,k} \cap \mathcal{X} \subseteq \{x \in \mathcal{X} : a\tau(x) < 0\}$.

In Lemma 5, statement (i) means that we cannot identify the optimal arm on the randomization region $\mathcal{R}_k$. Statement (ii) says that pulling arm a on the exploitation region $\cup_{j=1}^k \mathcal{E}_{a,j}$ is optimal. Statement (iii) shows that the support of the sample $S_{a,k}$ (i.e., $((\cup_{j=1}^k \mathcal{E}_{a,j}) \cup \mathcal{R}_k) \cap \mathcal{X}$) always contains the region in which arm a is optimal, $\mathcal{Q}_a$. Statement (iv) says that the optimal arm on $\mathcal{D}_{a,k}$ is $-a$, which justifies why we put $\mathcal{D}_{a,k}$ into $\mathcal{E}_{-a,k}$ in Equation (8). Recall that, on $\mathcal{D}_{a,k}$, the support of the sample $S_{a,k}$ is insufficiently regular, and thus, we cannot hope to obtain good estimates there. Fortunately, statement (iv) guarantees that accurate decision making is still *possible* on $\mathcal{D}_{a,k}$ even though accurate CATE estimation is *impossible*.

Statement (iii) in Lemma 5 is crucial. On the one hand, it is critical in guaranteeing that sufficient samples can be collected for both arms for future epochs (see also the discussion following Theorem 1). On the other hand, it leads to statement (iv), which enables us to make correct decisions in the inestimable regions. The argument is roughly as follows. Given statement (iii), if statement (iv) didn't hold, that is, if there were any $x_0 \in \mathcal{D}_{a,k} \cap \mathcal{X}$ such that $x_0 \in \mathcal{Q}_a = \{x \in \mathcal{X} : a\tau(x) \ge 0\}$, then by the strong regularity of $\mathcal{Q}_a$ imposed by Assumption 2, $((\cup_{j=1}^k \mathcal{E}_{a,j}) \cup \mathcal{R}_k) \cap \mathcal{X}$ would be sufficiently regular at $g(x_0)$, which violates the construction of $\mathcal{D}_{a,k}$ in Equation (6).

4.1.2. A Preliminary Regret Analysis.

Based on Lemma 5, we can decompose the regret according to $\overline{\mathcal{G}}_{k-1} \cap \overline{\mathcal{M}}_{k-1}$. Let $\hat{\pi}$ denote our algorithm, Algorithm 1. Then,

$$\begin{aligned} R_T(\hat{\pi}) &= \sum_{k=1}^K \sum_{t \in \mathcal{T}_k} \mathbb{E}[Y_t(\pi^*(X_t)) - Y_t(A_t)] \\ &\le \sum_{k=1}^K \sum_{t \in \mathcal{T}_k} \mathbb{E}[Y_t(\pi^*(X_t)) - Y_t(A_t) | \overline{\mathcal{G}}_{k-1} \cap \overline{\mathcal{M}}_{k-1}] \\ &\quad + \sum_{k=1}^K \sum_{t \in \mathcal{T}_k} \mathbb{P}\left(\overline{\mathcal{G}}_{k-1}^C \cup \overline{\mathcal{M}}_{k-1}^C \right). \end{aligned}$$

We can further decompose the regret in the k^{th} epoch given $\overline{\mathcal{G}}_{k-1} \cap \overline{\mathcal{M}}_{k-1}$ into the regret resulting from exploitation in

$\bigcup_{j=1}^{k}\mathcal{E}_{1,j}\cup\mathcal{E}_{-1,j}$ and the regret resulting from exploration in $\mathcal{R}_k$:

$$\begin{aligned}&\sum_{t\in\mathcal{T}_k}\mathbb{E}[Y_t(\pi^*(X_t))-Y_t(A_t)|\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1}]\\&\quad\le\sum_{t\in\mathcal{T}_k}\mathbb{E}\left[Y_t(\pi^*(X_t))-Y_t(A_t)|\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1},\right.\\&\qquad\left.X_t\in\left(\bigcup_{j=1}^{k}\mathcal{E}_{1,j}\bigcup\mathcal{E}_{-1,j}\right)\right]\\&\quad+\sum_{t\in\mathcal{T}_k}\mathbb{E}[|\tau(X_t)||\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1},X_t\in\mathcal{R}_k]\\&\qquad\times\mathbb{P}(X_t\in\mathcal{R}_k|\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1}).\end{aligned}$$

Lemma 5 statement (ii) implies that the proposed algorithm always pulls the optimal arm on the exploitation region. Therefore, the first term on the right-hand side, that is, the regret resulting from exploitation, is equal to zero. Moreover,

$$\begin{aligned}&\sum_{t\in\mathcal{T}_k}\mathbb{E}[|\tau(X_t)||\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1},X_t\in\mathcal{R}_k]\\&\quad\times\mathbb{P}(X_t\in\mathcal{R}_k|\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1})\\&\qquad\le\sum_{t\in\mathcal{T}_k}2\epsilon_{k-1}\mathbb{P}(0<|\tau(X_t)|\le2\epsilon_{k-1}|\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1})\\&\qquad\le\ \gamma2^{1+\alpha}\epsilon_{k-1}^{1+\alpha}n_k,\end{aligned}$$

where the first inequality follows from Lemma 5, statement (i), and the second inequality follows from the margin condition of Assumption 4.

Therefore, the total regret is bounded as follows:

$$\begin{aligned}R_T(\hat{\pi})&\le\sum_{k=1}^{K}\gamma2^{1+\alpha}\epsilon_{k-1}^{1+\alpha}n_k+\sum_{k=1}^{K}n_k\mathbb{P}\left(\overline{\mathcal{G}}_{k-1}^{C}\cup\overline{\mathcal{M}}_{k-1}^{C}\right)\\&\le O\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}\log^{1+\frac{d}{2\beta}}(T)+\log^2(T)\right)+\sum_{k=1}^{K}n_k\mathbb{P}\left(\overline{\mathcal{G}}_{k-1}^{C}\cup\overline{\mathcal{M}}_{k-1}^{C}\right),\end{aligned}\tag{13}$$

where the $O(\cdot)$ term depends only on the parameters of Assumptions 1–4 and not on the particular instance. Thus, if we can prove that $\mathbb{P}(\overline{\mathcal{G}}_{k-1}^{C}\cup\overline{\mathcal{M}}_{k-1}^{C})$ is small enough for all k, then we can (uniformly) bound the cumulative regret $R_T(\hat{\pi})$ of our proposed algorithm.

4.1.3. Bounding $\mathbb{P}(\overline{\mathcal{G}}_{k-1}^{C}\cup\overline{\mathcal{M}}_{k-1}^{C})$.

The analysis in Equation (13) shows that the cumulative regret of the proposed algorithm depends on the probability of $\overline{\mathcal{G}}_{k-1}^{C}\cup\overline{\mathcal{M}}_{k-1}^{C}$, that is, that the CATE estimator may not be accurate enough or that the total sample size for one arm is not sufficient in any epoch prior to the k^{th} epoch.

To bound this probability, we need to analyze the distribution of the samples for each arm. The sample distributions in each epoch can be distorted by decisions in previous epochs. Because a well-behaved density is crucial for nonparametric estimation, we must make sure that such distortions do not undermine our CATE estimation.

Lemma 6. *For any $1\le k\le K$ and $a=\pm1$, $S_{a,k}=\{(X_t,Y_t):t\in\mathcal{T}_k,A_t=a\}$ are conditionally i.i.d. samples, given $\mathcal{F}_{k-1}\cup\mathcal{A}_k$, where $\mathcal{F}_{k-1}=\{(X_t,A_t,Y_t):t\in\bigcup_{k'=1}^{k-1}\mathcal{T}_{k'}\}$, $\mathcal{A}_k=\{A_t:t\in\mathcal{T}_k\}$.*

Now suppose Assumptions 2 and 3 hold, let C_0 be defined as in Lemma 7 for any given $\beta,L_1,c_0,r_0,\mu_{\min}$ and suppose $T\ge T_0\vee\left(\exp\left(1\vee\frac{C_0(2\beta+d)}{4(2r_0)^{2\beta}(2\beta+d+\beta d)}\right)\right)$. Then, for $a=\pm1$ under the event $\overline{\mathcal{G}}_{k-1}\cap\overline{\mathcal{M}}_{k-1}$, the (common) conditional density of any of $\{X_t:A_t=a,t\in\mathcal{T}_k\}$ with respect to Lebesgue measure, given $\mathcal{F}_{k-1}\cup\mathcal{A}_k$, which we denote by $\mu_{a,k}$, satisfies the following conditions:

1. $\frac{1}{2}\mu_{\min}\le\mu_{a,k}(x)\le\frac{2\mu_{\max}}{p}$ *for any* $x\in((\bigcup_{j=1}^{k}\mathcal{E}_{a,j})\cup\mathcal{R}_k)\cap\mathcal{X}$.
2. $\mu_{a,k}(x)=0$ *for any* $x\in(\bigcup_{j=1}^{k}\mathcal{E}_{-a,j})\cap\mathcal{X}$.

Lemma 6 shows that, in the k^{th} epoch, samples for each arm are i.i.d. given the history, and it satisfies a strong density condition on the support of each sample, $((\bigcup_{j=1}^{k}\mathcal{E}_{a,j})\cup\mathcal{R}_k)\cap\mathcal{X}$. Furthermore, this distribution support set is sufficiently regular with respect to points in $\mathcal{R}_k\cap\mathcal{D}_{1,k}^{C}\cap\mathcal{D}_{-1,k}^{C}$, according to the screening rule given in Equation (6). Together, this strong density condition and support set strong regularity condition guarantee that we can estimate CATE using local polynomial estimators well on $\mathcal{R}_k$ in the $(k+1)^{\text{th}}$ epoch after we remove the inestimable regions.

In particular, the following lemma shows that the local polynomial estimator is well-conditioned with high probability, which echoes the classic result in the off-line setting (Proposition 3).

Lemma 7. *Suppose the conditions of Lemma 6 hold. Let $1\le k\le K-1,a=\pm1,n_{\pm1,k}$ be given. Consider the Gram matrices of the local polynomial regression estimators in Equation (11), that is, $\hat{\mathcal{A}}(x;S_{a,k},H_{a,k},\mathfrak{b}(\beta))$ as defined in Equation (5). Then, given $N_{\pm1,k}=n_{\pm1,k}$ and $\overline{\mathrm{M}}_{k-1}\cap\overline{\mathrm{G}}_{k-1}$, these satisfy the following with conditional probability at least $1-2M_\beta^2\exp\{-C_0(4(1+L_1\sqrt{d})^2)n_{a,k}^{2\beta/(2\beta+d)}\}$:*

$$\lambda_{\min}(\hat{\mathcal{A}}(x;S_{a,k},H_{a,k},\mathfrak{b}(\beta)))\ge\lambda_0>0,$$

$$\forall x\in\mathcal{R}_k\cap\mathcal{D}_{1,k}^{C}\cap\mathcal{D}_{-1,k}^{C},$$

where

$$\lambda_0=\frac{1}{4}\mu_{\min}\inf_{\substack{W\in\mathbb{R}^d,\,S\subset\mathbb{R}^d:\,\|W\|=1\\S\subseteq\mathcal{B}(0,1)\text{ is compact},\,\mathrm{Leb}(S)=c_0v_d/2^d}}\int_S\left(\sum_{|s|\le\mathfrak{b}(\beta)}W_su^s\right)^2du,$$

$$\begin{aligned}C_0=&\frac{3p\lambda_0^2}{4(1+L_1\sqrt{d})^2}\\&\times\min\left\{\frac{1}{12M_\beta^4\mu_{\max}v_d+2p\lambda_0M_\beta^2},\right.\\&\frac{1}{108M_\beta v_d\mu_{\max}+6\sqrt{M_\beta}p\lambda_0},\\&\left.\frac{1}{108M_\beta L^2v_d\mu_{\max}+6\sqrt{M_\beta}L(2v_d\mu_{\max}+p)\lambda_0}\right\}.\end{aligned}$$

In Lemma 7, λ_0 is positive because the unit shell is compact, and for fixed W, the infimum over S is continuous in W and positive as the integrand can be zero only in a measure-zero set, whereas S has a positive measure. The constant C_0 dictates the epoch schedule $\{\mathcal{T}_k\}_{k=1}^K$ of our proposed algorithm (see Section 3). Note that we can also use any positive constant no larger than C_0 in our algorithm without deteriorating the regret rate.

In the following theorem, we show that $\mathbb{P}(\overline{\mathcal{G}}_{k-1}^C \cup \overline{\mathcal{M}}_{k-1}^C)$ is indeed very small for large T, so its contribution to the cumulative regret bound in Equation (13) is negligible.

Theorem 1. *When* $T \geq T_0 \vee \left(\exp(1 \vee \frac{C_0(2\beta+d)}{4(2r_0)^{2\beta}(2\beta+d+\beta d)} \vee \frac{36M_\beta L^2 v_d^2 \mu_{\max}^2 C_0(2\beta+d)}{p^2\lambda_0^2(2\beta+d+\beta d)})\right)$, *if we assume Assumptions 1–3, then for any* $1 \leq k \leq K-1$,

$$\mathbb{P}(\mathcal{G}_k^C|\overline{\mathcal{G}}_{k-1}, \overline{\mathcal{M}}_k) \leq \frac{8+4M_\beta^2}{T},$$
$$\mathbb{P}(\mathcal{M}_k^C|\overline{\mathcal{G}}_{k-1}, \overline{\mathcal{M}}_{k-1}) \leq \frac{2}{T},$$
$$\mathbb{P}(\overline{\mathcal{G}}_k^C \cup \overline{\mathcal{M}}_k^C) \leq \frac{(10+4M_\beta^2)k}{T}.$$

Here, the upper bound on $\mathbb{P}(\mathcal{G}_k^C|\overline{\mathcal{G}}_{k-1}, \overline{\mathcal{M}}_k)$ is derived from the uniform convergence of local polynomial regression estimators (Stone 1982) given well-conditioned Gram matrices (which we ensure in Lemma 7) and sufficiently many samples for each arm (ensured by $\mathcal{M}_k$) whose sample distribution satisfies strong density condition (which we ensure in Lemma 6). The upper bound on $\mathbb{P}(\mathcal{M}_k^C|\overline{\mathcal{G}}_{k-1}, \overline{\mathcal{M}}_{k-1})$ arises from Lemmas 1 and 5, statement (iii), because they imply that $\mathbb{P}(X \in (\cup_{j=1}^k \mathcal{E}_{a,j}) \cup \mathcal{R}_k) \geq \mathbb{P}(X \in \mathcal{Q}_a) \geq p$ for $a = \pm 1$. As a result, at least a constant fraction of n_k many samples accumulates for each arm so that $\mathcal{M}_k$ holds with high probability as the n_k proposed in Equation (10) is sufficiently large. The upper bound on $\mathbb{P}(\overline{\mathcal{G}}_k^C \cup \overline{\mathcal{M}}_k^C)$ follows from the first two upper bounds by induction.

4.1.4. Regret Upper Bound.

Given Theorem 1 and Equation (13), we are now prepared to derive the final upper bound on our regret.

Theorem 2. *Suppose Assumptions 1–4 hold. Then,*

$$R_T(\pi) = O\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}\log^{1+\frac{d}{2\beta}}(T) + \log^2(T)\right) = \tilde{O}\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}} + 1\right),$$

where the $O(\cdot)$ *and* $\tilde{O}(\cdot)$ *terms only depend on the parameters of Assumptions 1–4. (An explicit form is given in the proof.)*

Proof sketch. Theorem 1 states that, for $2 \leq k \leq K$,

$$n_k\mathbb{P}\left(\overline{\mathcal{G}}_{k-1}^C \cup \overline{\mathcal{M}}_{k-1}^C\right) \leq n_k\frac{(10+4M_\beta^2)(k-1)}{T} \leq (10+4M_\beta^2)(k-1).$$

Furthermore, Lemma 2 implies that

$$K \leq \left\lceil \frac{\beta}{(2\beta+d)\log 2}\log T \right\rceil.$$

Thus,

$$\sum_{k=1}^K n_k\mathbb{P}(\overline{\mathcal{G}}_{k-1}^C \cup \overline{\mathcal{M}}_{k-1}^C) \leq (5+2M_\beta^2)K^2 \leq (5+4M_\beta^2)\frac{\beta^2\log^2 T}{(2\beta+d)^2\log^2 2} = \tilde{O}(1).$$

The final conclusion follows from Equation (13).

A complete and detailed proof is given in the online supplement. As noted at the start of Section 3.2, whereas our Algorithm 1 takes T as an input, we can obtain the same result as Theorem 2 for an algorithm that does not know T by simply calling Algorithm 1 with doubling horizons. Notice also that, although Algorithm 1 takes c_0 as input, the *rate* in T of the regret bound does not depend on it. Examining the explicit form of the regret bound reveals a polynomial dependence in the constant. Moreover, Assumption 2 with some $c_0 > 0$ always implies the same with any $c_0' \in (0, c_0]$. Therefore, if Assumption 2 holds so that some positive but unknown c_0 exists, then for T large enough, we can always use $1/\log(T)$ as input to our algorithm and still obtain the same regret rate up to polylogarithmic factors. If Assumption 2 does not hold, then Theorem 4 shows that the minimax regret is of a different order of magnitude altogether.

Because Algorithm 1 is an admissible policy, this yields an upper bound on the minimax regret.

Corollary 1. *Let any problem parameters be given. Then, for the corresponding class of contextual bandit problems* $\mathcal{P}$, *the minimax regret satisfies*

$$\mathcal{R}_T = \tilde{O}\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}\right).$$

4.2. Regret Lower Bound

In this section, we prove a matching lower bound (up to polylogarithmic factors) for the regret rate in Theorem 2 in the regime in which $\alpha\beta \leq d$. This means that there does not exist any other algorithm that can achieve a lower rate of regret for all smooth bandit instances in a given smoothness class. Thus, our algorithm achieves the minimax-optimal regret rate.

Theorem 3 (Regret Lower Bound). *Fix any positive parameters α, β, d, L, L_1 satisfying $\alpha\beta \le d$. For any admissible policy π and T, there exists a contextual bandit instance satisfying Assumptions 1–4 with the provided parameters such that*

$$\sup_{\mathbb{P}\in\mathcal{P}} R_T(\pi) = \Omega\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}\right), \tag{14}$$

where the $\Omega(\cdot)$ term only depends on the parameters of the class $\mathcal{P}$ and not on π. Hence, we also have $\mathcal{R}_T = \Omega\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}\right)$.

Proof sketch. Define the *inferior sampling rate* of a given policy π as the expected number of times that π disagrees with the oracle policy π^* (for a given instance $\mathbb{P}$), that is,

$$I_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(\pi^*(X_t) \neq \pi_t(X_t)\right)\right].$$

Lemma 3.1 in Rigollet and Zeevi (2010) relates $R_T(\pi)$ to $I_T(\pi)$: under Assumption 4,

$$I_T(\pi) = O(T^{\frac{1}{1+\alpha}} R_T(\pi)^{\frac{\alpha}{1+\alpha}}). \tag{15}$$

Note the implicit dependence of $I_T(\pi), R_T(\pi)$ on the instance $\mathbb{P}$.

We then construct a finite class, $\mathcal{H}$, of contextual bandit instances with smooth expected rewards and show, first, that $\mathcal{H} \subseteq \mathcal{P}$, that is, that our construction fits the provided parameters (in particular, our construction is fundamentally different from that in Rigollet and Zeevi (2010) as their construction approach is only suitable for nondifferentiable functions) and, second, that

$$\sup_{\mathbb{P}\in\mathcal{P}} I_T(\pi) \ge \sup_{\mathbb{P}\in\mathcal{H}} I_T(\pi) \ge \frac{1}{|\mathcal{H}|}\sum_{\mathbb{P}\in\mathcal{H}} I_T(\pi) = \Omega(T^{1-\frac{\alpha\beta}{2\beta+d}}). \tag{16}$$

We arrive at the final conclusion by combining Equations (15) and (16). □

A complete and detailed proof is given in the online supplement.

Note that, in Theorem 3, we allow α, β, d, L, L_1 to be given. The proof then constructs an example with appropriate values for the rest of the parameters, $c_0, r_0, \mu_{\max}, \mu_{\min}, \gamma$, for which the class of bandit problems $\mathcal{P}$ satisfies the preceding lower bound. This shows that the rate given in Theorem 2 is tight (for the regime $\alpha\beta \le d$).

We can, furthermore, show that our nonstandard Assumption 2 is necessary for achieving the minimax regret rate $\tilde{\Theta}\left(T^{\frac{\beta+d-\alpha\beta}{2\beta+d}} + 1\right)$. In particular, we prove that there exists a class of problem instances satisfying Assumptions 1, 3, and 4 but not necessarily Assumption 2 such that the corresponding regret lower bound is higher in order.

Theorem 4 (Regret Lower Bound Without Assumption 2). *Fix any positive parameters α, β, d, L, L_1. For any admissible policy π and T, and any constant $\Delta \in (\frac{\beta d}{\alpha\beta+d}, \beta \wedge \frac{d}{\alpha})$, there exists a contextual bandit instance satisfying Assumptions 1, 3, and 4 (but not necessarily Assumption 2) with the provided parameters such that*

$$\frac{R_T(\pi)}{T^{\frac{\beta+d-\alpha\beta}{2\beta+d}}} = \Omega\left(T^{\frac{d(\alpha+1)(\beta-\Delta)}{(2\Delta+d)(2\beta+d)}}\right) \to \infty \quad \text{as} \quad T \to \infty. \tag{17}$$

Proof sketch. Similar to the proof of Theorem 3, for any $\Delta \in (\frac{\beta d}{\alpha\beta+d}, \beta \wedge \frac{d}{\alpha})$, we can construct another finite class $\mathcal{H}'$ of bandit problems that satisfy Assumptions 1, 3, and 4 but not Assumption 2 with the provided parameters and show that

$$\sup_{\mathbb{P}\in\mathcal{H}'} I_T(\pi) \ge \frac{1}{|\mathcal{H}'|}\sum_{\mathbb{P}\in\mathcal{H}'} I_T(\pi) = \Omega(T^{1-\frac{\alpha\Delta}{2\Delta+d}}). \tag{18}$$

In conclusion, Equation (17) then follows from Equations (15) and (18).

We want to remark here that the constructed instances in the proof of Theorem 4 are only some special classes of irregular problems and need not be the worst ones, so even when taking $\Delta \to \frac{\beta d}{\alpha\beta+d}$ approaching the worst irregularity in Theorem 4, we do not believe the obtained rate is the right minimax rate under only Assumptions 1, 3, and 4. In fact, irregular support can be arbitrarily complicated, and to the best of our knowledge, there is no previous work even in the off-line regression setting that studies the minimax rates (estimation or regret) for a Hölder function class without regularity (even showing that it is strictly worse in the off-line regression setting appears new, which our example in Theorem 4 shows). Characterizing the optimal minimax rates without regularity for either off-line regression or online decision making is beyond our scope and remains a possible avenue of future investigation. The main purpose of Theorem 4 is only to show the necessity of Assumption 2 in Theorem 2 and to shed more light on the complexity of these irregular problems.

5. Practical Implementation and Numerical Investigation

Our primary focus is theoretically understanding the learnability of decision making in nonparametric settings by characterizing the minimax regret. Our proposed algorithms (Algorithm 1 and Algorithm 2 in the Online Appendix) are, therefore, intended primarily as an exhibited valid policy used to obtain a theoretically rigorous upper bound on the minimax regret for the smooth contextual bandit problem and not as a practically viable bandit algorithm. In this section, we use the primary insights from our algorithms to propose a simple, practical algorithm, and we use it for a simple numerical study.

5.1. The Simplified Algorithm

In this section, we use our proposed algorithm as an inspiration for a simplified algorithm. Several features of our algorithm were designed purely with its analysis in mind. Our algorithm is essentially an upper confidence approach wherein we explore when conditional mean arm rewards are indistinguishable from sufficient confidence and otherwise exploit the apparently better arm. The additional idiosyncrasies of our algorithm are necessary to make the approach amenable to analysis in our complex, nonparametric setting. These can probably be simplified or removed in practice, leaving only the primary upper confidence structure. Of course, this breaks our analysis, so the motivations for these simplifications are purely heuristic.

One such possibly extraneous feature is the algorithm's epoch structure; rather than proceeding in geometrically increasing epochs, in practice, one might choose to simply continually update one's estimates and confidence intervals at every round. Another is our piecewise constant grid structure, which was employed primarily to obtain uniform confidence intervals on mean reward functions via the union bound; in practice, one might simply produce local polynomial estimates at any new context that arrives. Related to this are our exact high-probability confidence intervals based on the piecewise constant structure of our estimates and upon which we decide whether to explore or exploit; likely any reasonable confidence interval works even with only approximate validity as long as we set the confidence appropriately. Finally, we believe that, in practice, one can probably safely abandon the screening of inestimable regions.

What remains are the primary features of our algorithm: when a new context arrives, use the data available to estimate the conditional mean arm rewards as well as confidence intervals on these, pull the seemingly better arm if the intervals do not intersect, and otherwise explore. (Or, in the multiarm setting, uniformly explore all arms that are not, thus, confidently dominated, which may only leave one arm; see Online Appendix A.) The insight from our algorithms suggests that the estimation step can be done using local polynomial regression of order $\mathfrak{b}(\beta)$ with bandwidth $c_1 t^{-1/(2\beta+d)}$ for some $c_1 > 0$ and that the confidence probabilities should be summable. Practically, for concreteness, we suggest using the standard two-sided confidence interval for ordinary least squares (which is an approximate confidence interval) fit on the polynomial features of the data within the bandwidth, and we suggest a confidence level of order c_2/t (albeit not summable just barely; one could additionally divide by $\log t$). Specifically, letting x denote the current context and $\varphi(x) \in \mathbb{R}^{M_\beta}$ its expansion into monomials of degree at most $\mathfrak{b}(\beta)$, $\hat{\mathcal{A}}_a(x)$ denote the Gram matrices (see also Equation (5)) for estimating each arm, $a = \pm 1$, $s_a(x)$ denote the observed averaged squared residuals, and ζ denote the $1 - t/(2c_2)$ quantile of the standard normal, we uniformly explore each arm whenever the difference of conditional mean reward estimates is within $\zeta\sqrt{\varphi(x)^\top\left(s_{+1}^2(x)\hat{\mathcal{A}}_{+1}^{-1}(x) + s_{-1}^2(x)\hat{\mathcal{A}}_{-1}^{-1}(x)\right)\varphi(x)}$ of zero and otherwise pull the better-seeming arm. To ensure good confidence intervals, if there are fewer than $2M_\beta$ data points from arm a within the bandwidth, then we pull arm a in this round. An algorithmic listing of the pseudocode for this procedure is provided in Algorithm 3 in Online Appendix C. Additionally, an implementation and replication code is available at https://github.com/CausalML/SmoothBandit.

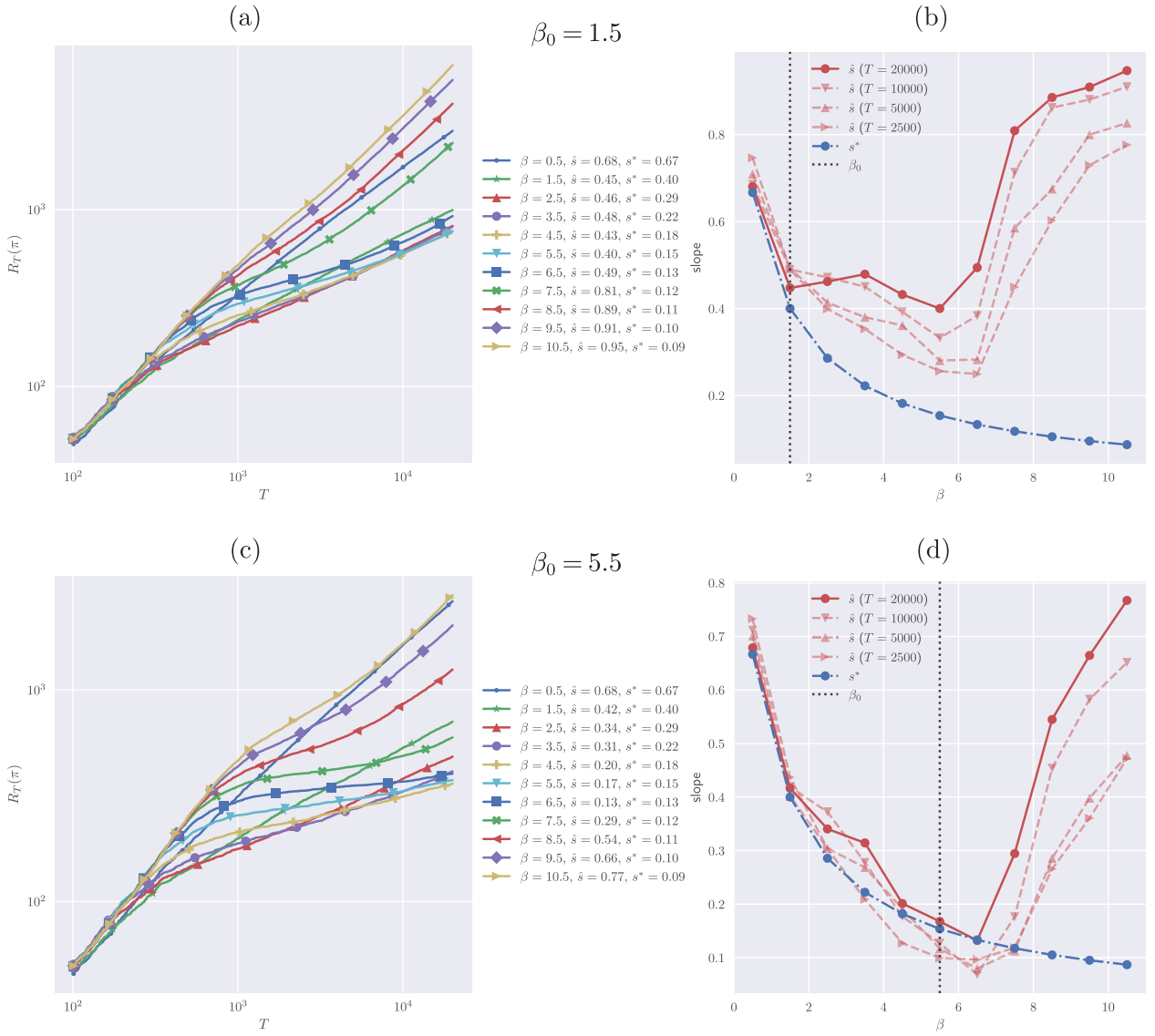
5.2. Numerical Study

In this section, we use our simplified algorithm to numerically study the smooth bandit problem. We consider covariates X drawn uniformly from the unit cube, $[0,1]^d$. And we consider random instances of the smooth bandit with η_{+1}, η_{-1} drawn independently from the Gaussian process prior on the unit cube with a Matérn covariance kernel with smoothness parameter β_0 and length-scale parameter 0.15 (the latter to ensure enough nonlinear behavior inside the unit cube). These random functions are β_0-smooth with probability one. We then consider running our simplified algorithm with $c_1 = 1, c_2 = 1/2$ and varying β.

In Figure 5, we present the regrets for every $T \in [1, \dots, 20000]$ averaged over five sample paths, each from a different random instance with $d = 2$, $\beta_0 \in \{1.5, 5.5\}$. For each sample path, we consider running our simplified algorithm with $\beta \in \{0.5, 1.5, \dots, 10.5\}$. The regrets are shown on a log–log scale. In addition to the regrets, for each algorithm, we report the slope $\hat{s}$ of the regret curve fit by least squared on the log–log data for $T \in [10000, \dots, 20000]$ as well as the theoretical asymptotic regret exponent $s^* = d/(2\beta + d)$ corresponding to the minimax regret rate exponent for the smooth bandit problem if smoothness is exactly β and the margin parameter is $\alpha = 1$ because, generically, we expect $\alpha = 1$ as long as contexts are continuous with a bounded density and η_a has an almost-everywhere nonsingular Jacobian. For comparison, we also include the slope $\hat{s}$ computed only on $T \in [T_0/2, \dots, T_0]$ for $T_0 = 2500, 5000, 10000$.

There are several observations to highlight. For $\beta \le \beta_0$, when the algorithm's smoothness parameter is well-specified, we note that the estimated and theoretical slopes match closely. First, this means that our minimax theory reliably predicts the regret behavior in practice. Second, this means that our simplified algorithm appears to achieve the correct regret rate. For β

Figure 5. (Color online) The Regret of Our Simplified Algorithm Using Different Smoothness Parameters β in Instances with Different True Smoothness β_0



Notes. Here, $\hat{s}$ refers to the slope of the log–log plots fitted to $T \in [10000, \ldots, 20000]$ and $s^* = d/(2\beta + d)$. (a) Regret of using different β under $\beta_0 = 1.5$. (b) Estimated and theoretical slopes; $\beta_0 = 1.5$. (c) Regret of using different β under $\beta_0 = 5.5$. (d) Estimated and theoretical slopes; $\beta_0 = 5.5$.

much bigger than β_0, when the algorithm's smoothness parameter is very badly specified, we see that the regret rate is worse than the theoretically predicted rate if β were well-specified. The estimated rate, however, appears sublinear only because T is finite. As T grows, the slope deteriorates toward one. This is also made clear by considering the apparent slopes we would have computed for shorter horizons: as we consider longer horizons, the slopes in the misspecified case get closer and closer to one. For β only slightly bigger than β_0, we surprisingly sometimes see a slope that is slightly better than $\beta = \beta_0$. This again, however, is only an effect of finite T. Even for very large β, we see an initially small slope that then inflects upward. The same happens for β slightly bigger than β_0 as we increase T; we eventually hit an inflection point at which the slope deteriorates toward one, and the inflection and deterioration would be slower for β very close to but above β_0. This, in fact, brings up an important practical point: our theoretical characterization of the minimax regret is only for T sufficiently large, in which case using the maximal correct smoothness, $\beta = \beta_0$, is optimal, but for shorter horizons T, there may be a benefit to slightly *oversmooth* $\beta > \beta_0$.

6. Conclusions

In this paper, we define and solve the smooth-response contextual bandit problem. We propose a rate-optimal algorithm that interpolates between using global and local reward information according to the underlying smoothness structure. Our results connect disparate results for contextual bandits and bridge the gap between linear-response and nondifferentiable bandits and contribute to revealing the whole landscape of contextual bandit regret and its interplay with the inherent complexity of the underlying learning problem.

Acknowledgments

The authors thank the anonymous review team for their very constructive comments. The authors are listed in alphabetical order.

References

Agarwal A, Hsu D, Kale S, Langford J, Li L, Schapire RE (2014) Taming the monster: A fast and simple algorithm for contextual bandits. *Proc. 31st Internat. Conf. Machine Learn.*, vol. 32 (JMLR.org), II–1638–II–1646.

Audibert JY, Tsybakov AB (2005) Fast learning rates for plug-in classifiers under the margin condition.

Audibert JY, Tsybakov AB (2007) Fast learning rates for plug-in classifiers. *Ann. Statist.* 35(2):608–633.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learn.* 47(2–3):235–256.

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (1995) Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proc. IEEE 36th Annual Foundations Comput. Sci.*, 322–331.

Bartlett PL, Bousquet O, Mendelson S (2005) Local Rademacher complexities. *Ann. Statist.* 33(4):1497–1537.

Bastani H, Bayati M (2020) Online decision making with high-dimensional covariates. *Oper. Res.* 68(1):276–294.

Bastani H, Bayati M, Khosravi K (2020) Mostly exploration-free algorithms for contextual bandits. *Management Sci.* 67(3):1329–1349.

Belloni A, Chernozhukov V, Chetverikov D, Kato K (2015) Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* 186(2):345–366.

Bertsimas D, Kallus N (2019) From predictive to prescriptive analytics. *Management Sci.* 66(3):1025–1044.

Beygelzimer A, Langford J, Li L, Reyzin L, Schapire R (2011) Contextual bandit algorithms with supervised learning guarantees. Gordon G, Dunson D, Dudík M, eds. *Proc. 14th Internat. Conf. Artificial Intelligence Statist., Proc. Machine Learn. Res.*, vol. 15 (PMLR, Fort Lauderdale, FL), 19–26.

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends® Machine Learn.* 5(1):1–122.

Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, Cambridge, UK).

Chen X (2007) Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, vol. 6, 5549–5632.

Dudik M, Hsu D, Kale S, Karampatziakis N, Langford J, Reyzin L, Zhang T (2011) Efficient optimal learning for contextual bandits. *Proc. 27th Conf. Uncertainty Artificial Intelligence* (AUAI Press, Arlington, VA), 169–178.

Fontaine X, Berthet Q, Perchet V (2019) Regularized contextual bandits. *The 22nd Internat. Conf. Artificial Intelligence Statist.*, 2144–2153.

Goldenshluger A, Zeevi A (2009) Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* 19(4):1603–1633.

Goldenshluger A, Zeevi A (2013) A linear response bandit problem. *Stochastic Systems* 3(1):230–261.

Gur Y, Momeni A, Wager S (2019) Smoothness-adaptive contextual bandits.

Lai T, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.

Langford J, Zhang T (2007) The epoch-greedy algorithm for contextual multi-armed bandits. *Proc. 20th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates Inc.), 817–824.

Mammen E, Tsybakov AB (1999) Smooth discrimination analysis. *Ann. Statist.* 27(6):1808–1829.

Nadaraya E (1964) On estimating regression. *Theory Probab. Appl.* 9(1):141–142.

Perchet V, Rigollet P (2013) The multi-armed bandit problem with covariates. *Ann. Statist.* 41(2):693–721.

Reeve H, Mellor J, Brown G (2018) The k-nearest neighbour UCB algorithm for multi-armed bandits with covariates. *Proc. Algorithmic Learn. Theory (PMLR)*, 83, 725–752.

Rigollet P, Zeevi A (2010) Nonparametric bandits with covariates. *Proc. 23rd Annual Conf. Learn. Theory*, 54–66.

Slivkins A (2011) Contextual bandits with similarity information. *Proc. 24th Annual Conf. Learn. Theory* (JMLR Workshop and Conference Proceedings), 679–702.

Steinwart I, Hush D, Scovel C (2006) An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory* 52(10):4635–4643.

Stone CJ (1977) Consistent nonparametric regression. *Ann. Statist.* 5(4):595–620.

Stone CJ (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* 8(6):1348–1360.

Stone CJ (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* 10(4):1040–1053.

Tsybakov AB (2004) Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* 32(1):135–166.

Tsybakov AB (2008) Introduction to nonparametric estimation. 1st ed. (Springer, New York).

Valko M, Korda N, Munos R, Flaounas I, Cristianini N (2013) Finite-time analysis of kernelised contextual bandits. *Proc. 29th Conf. Uncertainty Artificial Intelligence*, 654–663.

Wang CC, Kulkarni SR, Poor HV (2005) Bandit problems with side observations. *IEEE Trans. Automatic Control* 50(3): 338–355.

Watson GS (1964) Smooth regression analysis. *Sankhyā Indian J. Statist. Ser. A (1961–2002)* 26(4):359–372.

Yichun Hu is a PhD candidate in the School of Operations Research and Information Engineering at Cornell University. She is interested in various data-driven decision-making problems, especially in sequential settings.

Nathan Kallus is an assistant professor in the School of Operations Research and Information Engineering and Cornell Tech at Cornell University. Nathan's research interests include optimization, especially under uncertainty and informed by data; causal inference; sequential decision making; and algorithmic fairness.

Xiaojie Mao is an assistant professor in the Department of Management Science and Engineering at Tsinghua University. His research interests include data-driven decision making and causal inference.