Faster Rates in Regression Via Active Learning

Rui Castro, Rebecca Willett and Robert Nowak*

UW-Madison Technical Report ECE-05-3
rcastro@rice.edu, willett@cae.wisc.edu, nowak@engr.wisc.edu

June, 2005

Abstract

In this paper we address the theoretical capabilities of active sampling for estimating functions in noise. Specifically, the problem we consider is that of estimating a function from noisy point-wise samples, that is, the measurements which are collected at various points over the domain of the function. In the classical (passive) setting the sampling locations are chosen *a priori*, meaning that the choice of the sample locations precedes the gathering of the function observations. In the active sampling setting, on the other hand, the sample locations are chosen in an online fashion: the decision of where to sample next depends on all the observations made up to that point, in the spirit of the twenty questions game (as opposed to passive sampling where all the questions need to be asked before any answers are given). This extra degree of flexibility leads to improved signal reconstruction in comparison to the performance of classical (passive) methods. We present results characterizing the fundamental limits of active learning for various nonparametric function classes, as well as practical algorithms capable of exploiting the extra flexibility of the active setting and provably improving on classical techniques. In particular, significantly faster rates of convergence are achievable in cases involving functions whose complexity (in a the Kolmogorov sense) is highly concentrated in small regions of space (e.g., piecewise constant functions). Our active learning theory and methods show promise in a number of applications, including field estimation using wireless sensor networks and fault line detection.

1 Introduction

In this paper we address the theoretical capabilities of active learning for estimating functions in noise. In function regression, the goal is to estimate a

*Rui Castro is with the Department of Electrical and Computer Engineering, Rice University, Houston Texas. Robert Nowak and Rebecca Willett are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison Wisconsin. Corresponding Author: Rui Castro, E-mail: rcastro@rice.edu.

function from noisy point-wise samples. In the classical (passive) setting the sampling locations are chosen *a priori*, meaning that the selection of the sample locations precedes the gathering of the function observations. In the active sampling setting, however, the sample locations are chosen in an online fashion: the decision of where to sample next depends on all the observations made up to that point, in the spirit of the "Twenty Questions" game (as opposed to passive sampling, where all the questions need to be asked before any answers are given). The extra degree of flexibility garnered through active learning can lead to significantly better function estimates than those possible using classical (passive) methods. Several empirical and theoretical studies have shown that selecting samples or making strategic queries in order to learn a target function/classifier can outperform commonly used passive methods based on random or deterministic sampling [5, 16, 9, 18, 2]; however, there are very few analytical methodologies for these Twenty Questions problems when the answers are not entirely reliable (see for example [4, 10, 12]). This precludes performance guarantees and limits the applicability of many such methods in a theoretically sound way.

In the regression setting, significantly faster rates of convergence are achievable in cases involving functions whose complexity (in a the Kolmogorov sense) is highly concentrated in small regions of space (*e.g.*, functions that are smoothly varying apart from highly localized abrupt changes such as jumps or edges). We illustrate this by characterizing the fundamental limits of active learning for two broad nonparametric function classes which map $[0,1]^d$ onto the real line: Hölder smooth functions (spatially homogeneous complexity) and piecewise constant functions that are constant except on a $d-1$ dimensional *boundary set* or discontinuity embedded in the d dimensional function domain (spatially concentrated complexity). We conclude that, when the functions are spatially homogeneous and smooth, passive learning algorithms are near-minimax optimal over all estimation methods and all (active or passive) learning schemes, indicating that active learning methods will not lead to faster rates of convergence in this regime. For piecewise constant functions, active learning methods can capitalize on the highly localized nature of the boundary by focusing the sampling process in the estimated vicinity of the boundary. We present an algorithm that provably improves on the best possible passive learning algorithm and achieves faster rates of error convergence. Furthermore, we show that no other active learning method can significantly improve upon this performance (in a minimax sense).

Our active learning theory and methods show promise for a number of problems. In particular, in imaging techniques such as laser scanning it is possible to adaptively vary the scanning process. Using active learning in this context can significantly reduce image acquisition times. Wireless sensor networks constitute another key application area. Because of necessarily small batteries, it is desirable to limit the number of measurements collected as much as possible. Incorporating active learning strategies into such systems can dramatically lengthen the lifetime of the system. In fact, active learning problems like the one described in pose in Section 4.2 have already found application in fault line

detection [10] and boundary estimation in wireless sensor networking [20].

This paper is organized as follows: Section 2 describes the general scenario and framework. Section 3 describes the fundamental limits (in a minimax sense) of active and passive learning for various function classes. In Section 4 learning strategies are presented, both for passive and active learning. The performance of these strategies is also analyzed in the section. Finally Section 5 presents some concluding remarks and open problems and questions. The proofs of the results are presented in the Appendix.

2 Problem Statement

Let $\mathcal{F}$ denote a class of functions mapping $[0,1]^d$ to the real line. Later we will consider particular classes $\mathcal{F}$. Let $f : [0,1]^d \to \mathbb{R}$ be a function in that class. Our goal is to estimate this function from a finite number of noise corrupted samples. In this paper we consider two different scenarios: (a) *Passive learning*, where the locations of the sample points are chosen statistically independently way of the measurement outcomes. (b) *Active learning*, where the location of the i^{th} sample point can be chosen as a function of the samples points and samples collected up to that instant. The statistical model we consider builds on the following assumptions:

(A1) The observations $\{Y_i\}_{i=1}^n$ are given by

$$Y_i = f(\boldsymbol{X_i}) + W_i,\ i \in \{1, \ldots, n\}.$$

(A2) The random variables W_i are Gaussian zero mean and variance σ^2. These are independent and identically distributed (i.i.d.) and independent of $\{\boldsymbol{X_i}\}_{i=1}^n$.

(A3.1) Passive Learning: The sample locations $\boldsymbol{X_i} \in [0,1]^d$ are possibly random, but independent of $\{Y_i\}_{j \in \{1,\ldots,i-1,i+1,\ldots,n\}}$. They do not depend in any way on f.

(A3.2) Active Learning: The sample locations $\boldsymbol{X_i}$ are random, and depend only on $\{\boldsymbol{X_j}, Y_j\}_{j=1}^{i-1}$. In other words

$$\boldsymbol{X_i}|\boldsymbol{X_1} \ldots \boldsymbol{X_{i-1}}, \boldsymbol{X_{i+1}}, \ldots, \boldsymbol{X_n}, Y_1 \ldots Y_{i-1}, Y_{i+1}, \ldots, Y_n \overset{\text{a.s.}}{=} \boldsymbol{X_i}|\boldsymbol{X_1} \ldots \boldsymbol{X_{i-1}}, Y_1 \ldots Y_{i-1}.$$

Said in a different way, the sample locations $\boldsymbol{X_i}$ have only a causal dependency on the system variables $\{\boldsymbol{X_i}, Y_i\}$. Finally given $\{\boldsymbol{X_j}, Y_j\}_{j=1}^{i-1}$ the random variable $\boldsymbol{X_i}$ does not depend in any way on f.

Clearly the passive learning strategy is a special case of the active learning one. For the scenarios addressed in the paper we assume either (A3.1) or (A3.2) holds (respectively the passive and active learning scenarios).

The performance metric we consider is the usual L_2 norm,

$$d(f,g) \equiv \|f-g\| = \left(\int_{[0,1]^d} |f(\boldsymbol{x}) - g(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \right)^{1/2},$$

where $f, g \in \mathcal{F}$.

An estimator is a function $\hat{f}_{\boldsymbol{X_1},\ldots,\boldsymbol{X_n},Y_1\ldots,Y_n} : [0,1]^d \to \mathbb{R}$. That is, given $\{\boldsymbol{X_i}, Y_i\}_{i=1}^n$, $\hat{f}_{\boldsymbol{X_1},\ldots,\boldsymbol{X_n},Y_1\ldots,Y_n}(\cdot)$ is a function mapping $[0,1]^d$ to the real line. We will usually drop the explicit dependence of the estimator on $\{\boldsymbol{X_i}, Y_i\}_{i=1}^n$, and denote the estimator by $\hat{f}_n(\cdot)$.

When choosing an estimator $\hat{f}_n$ our main concern is to ensure that $d(f, \hat{f}_n)$ is small. In our model there is also another degree of freedom: we are allowed to choose our *sampling strategy*, that is, we can specify $\boldsymbol{X_i}|\boldsymbol{X_1}\ldots\boldsymbol{X_{i-1}}, Y_1\ldots Y_{i-1}$. We will denote the sampling strategy by S_n. A pair $(\hat{f}_n, S_n)$ is called a *estimation strategy*. In the rest of the paper we are going to study the fundamental performance limits for active learning for certain class of functions $\mathcal{F}$, and describe practical estimation strategies that nearly achieve those fundamental limits.

As discussed in Section 1, the extra degree of spatial adaptivity under (A3.2) can provide some gains when the functions in the class $\mathcal{F}$ have well localized features. In this case as the number of samples increases we can "focus" the sampling on the features that are impairing the estimation performance. Some classes we are going to consider in the paper have this property.

Assumption (A2) can be relaxed. Actually we only need the variables W_i to be independent, and their distribution needs to satisfy a certain "moment condition" (see the statement of Theorem 7). Many random variables satisfy that condition (*e.g.*, bounded random variables). To avoid cumbersome derivations we stick with the Gaussian assumption throughout the paper, although it is easy to generalize the results.

We are going to consider essentially two different types of functions: functions that are uniformly smooth; and functions that are piecewise constant/smooth, in the sense that these are comprised of constant/smooth regions, separated by boundaries that have upper box-counting dimension[1] at most $d-1$.

Definition 1. *A function $f : [0,1]^d \to \mathbb{R}$ is **locally Hölder smooth** at point $\boldsymbol{x} \in [0,1]^d$ if it has continuous partial derivatives up to order $k = \lfloor \alpha \rfloor$[2] at point $\boldsymbol{x} \in [0,1]^d$ and*

$$\exists \epsilon > 0 : \forall\ \boldsymbol{z} \in [0,1]^d : \quad \|\boldsymbol{z} - \boldsymbol{x}\| < \epsilon \ \Rightarrow\ |f(\boldsymbol{z}) - P_{\boldsymbol{x}}(\boldsymbol{z})| \leq L\|\boldsymbol{z} - \boldsymbol{x}\|^\alpha, \quad (1)$$

where $L, \alpha > 0$, and $P_{\boldsymbol{x}}(\cdot)$ denotes the degree k Taylor polynomial approximation of f expanded around $\boldsymbol{x}$.

[1] The upper box-counting dimension of a set B is defined using a cover of the set by closed balls of diameter r: Let $N(r)$ denote the minimal number of closed balls of diameter r that are cover of B, then the upper box-counting dimension of B is defined as $\limsup_{r\to 0} -\log N(r)/\log r$. The upper box-counting dimension is also known as the entropy dimension.

[2] $k = \lfloor \alpha \rfloor$ is the maximal integer such that $k < \alpha$.

If a function satisfies (1) *for every point* $\boldsymbol{x} \in [0,1]^d$ *the function is said to be* ***Hölder smooth*** *with parameters* L *and* α*. Denote this class of functions by* $\Sigma(L, \alpha)$*.*

Definition 2. *A function* $f : [0,1]^d \to \mathbb{R}$ *is* ***piecewise constant*** *if it is locally constant*[3] *in any point* $\boldsymbol{x} \in [0,1]^d \setminus B(f)$*, where* $B(f) \subseteq [0,1]^d$ *is a set with upper box-counting dimension at most* $d-1$*. Furthermore let* f *be uniformly bounded on* $[0,1]^d$ *(that is,* $|f(\boldsymbol{x})| \leq M$*,* $\forall \boldsymbol{x} \in [0,1]^d$*) and let* $B(f)$ *satisfy* $N(r) \leq \beta r^{-(d-1)}$ *for all* $r > 0$*, where* $\beta > 0$ *is a constant and* $N(r)$ *is the minimal number of closed balls of diameter* r *that covers* $B(f)$*. The set of all piecewise constant functions* f *satisfying the above conditions is denoted by* $PC(\beta, M)$*.*

Definition 3. *A function* $f : [0,1]^d \to \mathbb{R}$ *is* ***piecewise smooth*** *if* (1) *holds for any point* $\boldsymbol{x} \in [0,1]^d \setminus B(f)$*, where* $B(f) \subseteq [0,1]^d$ *is a set with upper box-counting dimension at most* $d-1$*. Furthermore let* f *be uniformly bounded on* $[0,1]^d$ *(that is,* $|f(\boldsymbol{x})| \leq M$*,* $\forall \boldsymbol{x} \in [0,1]^d$*) and let* $B(f)$ *satisfy* $N(r) \leq \beta r^{-(d-1)}$ *for all* $r > 0$*, where* $\beta > 0$ *is a constant and* $N(r)$ *is the minimal number of closed balls of diameter* r *that covers* $B(f)$*. The set of all piecewise smooth functions* f *satisfying the above conditions is denoted by* $PS(L, \alpha, \beta, M)$*.*

The concept of box-counting dimension is related the concept of topological dimension of a set [8], and these coincide when the set is "well-behaved". Essentially this condition means that the "boundaries" between the various smooth regions are $(d-1)$-dimensional non-fractal curves. We will frequently refer to the set $B(f)$ as the *boundary set*. The bound on $N(r)$ in the above definition leads to a bound on the upper-box counting dimension class. If the boundaries $B(f)$ are reasonably smooth then β is an approximate bound on the $d-1$ dimensional volume of $B(f)$.

The classes $\mathrm{PC}(\beta, M)$ and $\mathrm{PS}(L, \alpha, \beta, M)$ have the main ingredients that make active learning appealing: a function $f \in \mathrm{PS}(L, \alpha, \beta, M)$ is "well-behaved" everywhere in the unit square, except in the small set $B(f)$. We will see that the critical task for any estimator of f is accurately finding the location of the boundary $B(f)$.

3 Fundamental Limits - Minimax Lower Bounds

In this section we study the fundamental limitations of the active learning strategy. We start by introducing some notation.

Definition 4. *For any estimation strategy* $(\hat{f}_n, S_n)$*, and any element* $f \in \mathcal{F}$ *we define the* ***risk*** *of the estimation strategy as*

$$R(\hat{f}_n, S_n, f) = \mathbb{E}_{f,S_n}[d^2(\hat{f}_n, f)],$$

[3] A function $f : [0,1]^d \to \mathbb{R}$ is locally constant at a point $\boldsymbol{x} \in [0,1]^d$ if

$$\exists \epsilon > 0 \ : \forall \boldsymbol{y} \in [0,1]^d : \quad \|\boldsymbol{x} - \boldsymbol{y}\| < \epsilon \ \Rightarrow \ f(\boldsymbol{y}) = f(\boldsymbol{x}).$$

where $\mathbb{E}_{f,S_n}$ *is the expectation with respect to the probability measure of* $\{\boldsymbol{X_i}, Y_i\}_{i=1}^n$ *induced by model* f *and sampling strategy* S_n*. We define the* ***maximal risk*** *of an estimation strategy as* $\sup_{f\in\mathcal{F}} R(\hat{f}_n, S_n, f)$*.*

The goal of this section is to find tight lower bounds for the maximal risk, over all possible estimation strategies. That is, we present bounds of the form

$$\inf_{(\hat{f}_n,S_n)\in\Theta} \sup_{f\in\mathcal{F}} \mathbb{E}_{f,S_n}[d^2(\hat{f}_n, f)] \geq c\psi_n^2,\ \forall n \geq n_0 \tag{2}$$

where $n_0 \in \mathbb{N}$, $c > 0$ is a constant, ψ_n is a positive sequence converging to zero, and Θ is the set of all estimation strategies. The sequence ψ_n^2 is denoted as a *lower rate of convergence*[4].

It is also possible to devise upper bounds on the maximal risk. These are usually obtained through explicit estimation strategies (as presented in Section 4). If (2) and

$$\inf_{(\hat{f}_n,S_n)\in\Theta} \sup_{f\in\mathcal{F}} \mathbb{E}_{f,S_n}[d^2(\hat{f}_n, f)] \leq C\psi_n^2,\ \forall n \geq n_0 \tag{3}$$

hold, where $C > 0$, then ψ_n^2 is said to be the **optimal rate of convergence**. For the problems considered in this paper we observe that the rates are well approximated by $n^{-\gamma}$ for some γ, and a large n. When talking about optimal rates of convergence we are only interested in the polynomial behavior, therefore a rate of convergence ψ_n^2 is equivalent to $n^{-\gamma}$ (*i.e.*, $\psi_n^2 \overset{\text{poly}}{\sim} n^{-\gamma}$) if and only if given $\gamma_1 < \gamma < \gamma_2$ we have $n^{-\gamma_2} < \psi_n^2 < n^{-\gamma_1}$ for n large enough.

3.1 Passive Learning Minimax Rates

The passive learning model has been studied extensively for various classes $\mathcal{F}$, and there is a vast statistical literature on the optimal rates of convergence [17, 14].

For the class $\Sigma(L, \alpha)$ we have the following lower bound.

Theorem 1. *Under the requirements of the passive learning model we have*

$$\inf_{(\hat{f}_n,S_n)\in\Theta_{passive}} \sup_{f\in\Sigma(L,\alpha)} \mathbb{E}_{f,S_n}[\|\hat{f}_n - f\|^2] \geq cn^{-\frac{2\alpha}{2\alpha+d}}, \tag{4}$$

for n *large enough, where* $c \equiv c(L, \alpha, \sigma^2) > 0$*, and* $\Theta_{passive}$ *denotes the set of all passive estimation strategies.*

It is possible to show that the rate in the theorem is the optimal rate of convergence. A relatively simple linear estimator can achieve the above rate, using a deterministic sampling strategy. For a proof see for example [14]. If

[4]Clearly ψ_n^2 is defined up to a bounded factor, that depends on n. Namely, two rates of convergence ψ_n^2 and $\psi_n'^2$ are equivalent if and only if

$$0 < \lim\inf_{n\to\infty} \psi_n^2/\psi_n'^2 \leq \lim\sup_{n\to\infty} \psi_n^2/\psi_n'^2 < \infty.$$

we are concerned only with the rate up to polynomial equivalence then the estimator proposed in Section 4.1 achieves the desired rate.

We now turn our attention to the class of piecewise constant functions $\mathrm{PC}(\beta, M)$. A smaller class of functions, the boundary fragments, is studied in [14]. Let $g : [0,1]^{d-1} \to [0,1]$ be a Lipshitz function with graph in $[0,1]^d$, that is

$$g \in \left\{ b(\cdot) : |b(\boldsymbol{x}) - b(\boldsymbol{z})| \leq \|\boldsymbol{x} - \boldsymbol{z}\|,\ 0 \leq b(\boldsymbol{x}) \leq 1,\ \forall\ \boldsymbol{x}, \boldsymbol{z} \in [0,1]^{d-1} \right\}.$$

Define

$$G = \{(\boldsymbol{x}, y) : 0 \leq y \leq g(\boldsymbol{x}),\ \boldsymbol{x} \in [0,1]^{d-1}\}. \tag{5}$$

Finally define $f : [0,1]^d \to \mathbb{R}$ by $f(\boldsymbol{x}) = 2M\mathbf{1}_G(\boldsymbol{x}) - M$. The class of all the functions of this form is called the *boundary fragment* class (usually $M = 1$), denoted by $\mathrm{BF}(M)$. It is straightforward to to show that $\mathrm{BF}(M) \subseteq \mathrm{PC}(\beta, M)$, for a suitable constant β. Under the passive model we consider in this paper, we have the following result.

Theorem 2 (Korostelev-Tsybakov, 1993)**.** *Under the requirements of the passive learning model we have*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{passive}} \sup_{f \in BF(M)} \mathbb{E}_{f, S_n}[\|\hat{f}_n - f\|^2] \geq c n^{-\frac{1}{d}}, \tag{6}$$

for n large enough, where $c \equiv c(M, \sigma^2) > 0$.

It can be shown that the above bound is tight, in the sense that a corresponding upper-bound (3) holds and the rate in the theorem is the optimal rate of convergence. Noticing that $\mathrm{BF}(M) \subseteq \mathrm{PC}(\beta, M)$ we obtain

Proposition 1. *Under the requirements of the passive learning model we have*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{passive}} \sup_{f \in PC(\beta, M)} \mathbb{E}_{f, S_n}[\|\hat{f}_n - f\|^2] \geq c n^{-\frac{1}{d}}, \tag{7}$$

for n large enough, where $c \equiv c(\beta, M, \sigma^2) > 0$.

Also, since $\Sigma(L, \alpha) \subseteq \mathrm{PS}(L, \alpha, \beta, M)$ and $\mathrm{BF}(M) \subseteq \mathrm{PS}(L, \alpha, \beta, M)$ we can put together the results of Theorems 1 and 2 and get the following result for the piecewise smooth functions.

Proposition 2. *Under the requirements of the passive learning model we have*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{passive}} \sup_{f \in PS(L, \alpha, \beta, M)} \mathbb{E}_{f, S_n}[\|\hat{f}_n - f\|^2] \geq c \max\left\{ n^{-\frac{2\alpha}{2\alpha + d}}, n^{-\frac{1}{d}} \right\}, \tag{8}$$

for n large enough, where $c \equiv c(L, \alpha, \beta, M, \sigma^2) > 0$.

It is possible to construct estimators that achieve, or nearly achieve, the above performance rates, as we will see in Section 4.

Observe that, for α small enough (namely $\alpha < 1/(2-2/d)$), the lower bounds of Theorem 1 and Proposition 2 coincide. This indicates that in such cases the existence of a boundary set in the elements of $\mathrm{PS}(L,\alpha,\beta,M)$ is not really making the problem harder. On the other hand, when the function pieces are sufficiently smooth, the rate of convergence for $\mathrm{PS}(L,\alpha,\beta,M)$ is bounded below by $O(n^{-1/d})$. This corresponds to the contribution of the boundary, and that rate does not change even if we make α larger. This is in contrast to the behavior for $\Sigma(L,\alpha)$, where as α increases the performance rate gets arbitrarily close to the parametric rate $O(1/n)$. This indicates that the class of piecewise smooth functions may benefit from an active learning strategy as long as the function is reasonably smooth away from the boundary. Below we see that this is indeed the case.

3.2 Active Learning Minimax Rates

We begin by studying the class $\Sigma(L,\alpha)$ under the active learning model. Intuitively, there should be no advantage of active learning under this model class, since there are no localized features. Classical approximation theory results also support this intuition: the best m-term approximation scheme for the Hölder class of functions is a linear scheme, using a piecewise polynomial fit. We have the following main result.

Theorem 3. *Under the requirements of the active learning model we have*

$$\inf_{(\hat{f}_n,S_n)\in\Theta_{active}} \sup_{f\in\Sigma(L,\alpha)} \mathbb{E}_{f,S_n}[\|\hat{f}_n - f\|^2] \geq cn^{-\frac{2\alpha}{2\alpha+d}}, \tag{9}$$

for n large enough, where $c \equiv c(L,\alpha,\sigma^2) > 0$, where Θ_{active} is the set of all active estimation strategies.

Note that the rate in Theorem 3 is the same as the classical passive learning rate [17, 14] but the class of estimation strategies allowed is now much bigger. The proof of Theorem 3 is presented in Appendix A. There are various practical estimators achieving the performance predicted by Theorem 3, including some based on kernels, splines or wavelets [19].

We now turn our attention to the class of piecewise constant functions $\mathrm{PC}(L,\alpha,\beta,M)$. In [12, 13] the active learning scenario was studied for the boundary fragment class. For this class the following result holds.

Theorem 4 (Korostelev, 1999)**.** *Let $d \geq 2$. Under the requirements of the active learning model we have*

$$\inf_{(\hat{f}_n,S_n)\in\Theta_{active}} \sup_{f\in BF(M)} \mathbb{E}_{f,S_n}[\|\hat{f}_n - f\|^2] \geq cn^{-\frac{1}{d-1}}, \tag{10}$$

for n large enough, where $c \equiv c(M,\sigma^2) > 0$.

The above result is restricted to $d \geq 2$. For $d = 1$ we have an exponential rate of convergence (much faster than the passive parametric rate of $1/n$). This was shown in pioneering work of Burnashev and Zigangirov [4].

In [14] an algorithm capable of achieving the above rate for the boundary fragment class is presented, but this algorithm takes advantage of the very special functional form of the boundary fragment functions. The algorithm begins by dividing the unit hypercube into "strips" and performing a one-dimensional change-point estimation in each of the strips. This change-point detection can be performed extremely accurately using active learning, as shown in [4]. Unfortunately, the boundary fragment class is very restrictive and impractical for most applications. Recall that boundary fragments consist of only two regions, separated by a boundary that is a function of the first $d-1$ coordinates. The class $\mathrm{PC}(\beta, M)$ is much larger and more general, so the algorithmic ideas that work for boundary fragments can no longer be used. A completely different approach is required, using radically different tools.

From Theorem 4 we obtain one of the main results of this section.

Theorem 5 (Minimax Lower Bound for $\mathrm{PC}(\beta, M)$)**.** *Let $d \geq 2$. Under the requirements of the active learning model we have*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{active}} \sup_{f \in PC(\beta, M)} \mathbb{E}_{f, S_n}[\|\hat{f}_n - f\|^2] \geq c n^{-\frac{1}{d-1}}, \tag{11}$$

for n large enough, where $c \equiv c(\beta, M, \sigma^2) > 0$.

In contrast with Proposition 1, we observe that with active learning we have a potential performance gain over passive strategies, effectively equivalent to a dimensionality reduction. Essentially the exponent in (11) depends now on the dimension of the boundary set, $d-1$, instead of the dimension of the entire domain, d. In the next section we verify that the bound in Theorem 5 is tight, and present an algorithm whose performance is arbitrarily close to that bound (in terms of the polynomial rate).

Taking into account the results of Theorem 3 and Theorem 4 we obtain following result.

Theorem 6 (Minimax Lower Bound for $\mathrm{PS}(L, \alpha, \beta, M)$)**.** *Let $d \geq 2$. Under the requirements of the active learning model we have*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{active}} \sup_{f \in PS(L, \alpha, \beta, M)} \mathbb{E}_{f, S_n}[\|\hat{f}_n - f\|^2] \geq c \max\left\{ n^{-\frac{2\alpha}{2\alpha+d}}, n^{-\frac{1}{d-1}} \right\}, \tag{12}$$

for n large enough, where $c \equiv c(L, \alpha, \beta, M, \sigma^2) > 0$.

From this bound we see that active learning can benefit estimation strategies for the piecewise smooth class. We conjecture that the rate in the Theorem is actually the optimal rate, although we have not yet proved this. Therefore the answer to this question remains an open problem.

4 Learning Strategies

In this section we present various estimation strategies, both for the passive and active learning models. All the estimation strategies hinge on a tree structured partition, that allows for the necessary degree of spatial adaptivity. The

design and analysis of the proposed algorithms are intertwined, since we use various bounding techniques as guidelines in the design process. The following fundamental risk bound is a key tool.

Theorem 7. *Assume (A1) and (A3.1). Furthermore let* $\{\boldsymbol{X}_i\}_{i=1}^n$ *be i.i.d., uniform over* $[0,1]^d$*, and independent of* $\{Y_i\}_{i=1}^n$ *and* $\{\boldsymbol{X}_i\}_{i=1}^n$*. Suppose also that for all* $i \in \{1,\ldots,n\}$ *we have* $\mathbb{E}[W_i] = 0$*,* $Var(W_i) \leq \sigma^2$*, and*

$$\mathbb{E}\left[|W_i|^k\right] \leq Var(W_i)\frac{k!}{2}h^{k-2}, \tag{13}$$

for some $h > 0$ *and* $k \geq 2$*. Equation* (13) *is known as the* Bernstein's moment condition.

Let Γ *be a countable class of functions mapping* $[0,1]^d$ *to the real line such that*

$$|f(x)| \leq M \quad \forall x \in [0,1]^d, \forall f \in \Gamma.$$

Let $\mathrm{pen} : \Gamma \to [0,+\infty)$ *be a function satisfying*

$$\sum_{\theta' \in \Gamma} e^{-\mathrm{pen}(\theta')} \leq s, \tag{14}$$

for some $s > 0$*.*

Finally define the estimator

$$\hat{f}_n(\boldsymbol{X},\boldsymbol{Y}) \equiv \arg\min_{f' \in \Gamma}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - f'(\boldsymbol{X}_i))^2 + \frac{\lambda}{n}\mathrm{pen}(f')\right\}, \tag{15}$$

where $\lambda > 2(\sigma^2 + M^2) + 32(hM + M^2/3)$*.*

Then

$$\mathbb{E}\left[\|f - \hat{f}_n\|^2\right] \leq \min_{f' \in \Gamma} \frac{1}{1-a}\left\{(1+a)\|f - f'\|^2 + \frac{\lambda}{n}\mathrm{pen}(f') + \frac{2\lambda(s+1)}{n}\right\}, \tag{16}$$

with $a = \frac{2(\sigma^2+M^2)}{\lambda - 32(hM+M^2/3)}$*.*

This theorem is an oracle bound, that is, the expected error of the estimator is, up to a multiplicative constant, the best possible relative to the penalized criterion among all the models in Γ. The proof is presented in Appendix B. Note that there is some freedom when deciding how important the penalty is (parameter λ).

Remark: Notice that (14) can be interpreted as a Kraft inequality [6]. This means that we can construct the penalty function $\mathrm{pen}(\cdot)$ by explicitly describing a prefix code for the elements of Γ: for each $\theta \in \Gamma$, $\mathrm{pen}(\theta)$ is the length (in *nats*) of the codeword associated with θ. The coding argument is sometimes convenient since a prefix code automatically satisfies (14).

4.1 Passive Learning Strategies

In this section we focus on the passive learning model (A3.1) and the class of functions $\mathrm{PS}(L, \alpha, \beta, M)$. A simplification of the techniques employed allows us to get similar results for the piecewise constant functions in the class class $\mathrm{PC}(B, M)$.

Since the location of the boundary is *a priori* unknown, it is natural to distribute the sample points in a uniform way over the unit hypercube. Various sampling schemes can be used to accomplish this, but we focus on a very simple randomized scheme. Let $\{\boldsymbol{X}_i\}_{i=1}^n$ be i.i.d. uniform over $[0,1]^d$. Under assumption (A2) W_i is Gaussian and so we obtain the following corollary of Theorem 7.

Corollary 1. *Assume (A1), (A2) and (A3.1). Furthermore let $\{\boldsymbol{X}_i\}_{i=1}^n$ be i.i.d., uniform over $[0,1]^d$, and independent of $\{Y_i\}_{i=1}^n$ and $\{\boldsymbol{X}_i\}_{i=1}^n$. Consider a class of models Γ satisfying the conditions of Theorem 7 and define the estimator*

$$\hat{f}_n(\boldsymbol{X}, \boldsymbol{Y}) \equiv \arg\min_{f' \in \Gamma} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f'(\boldsymbol{X}_i))^2 + \frac{\lambda}{n} \mathrm{pen}(f') \right\}, \tag{17}$$

with $\lambda = 6(\sigma^2 + M^2) + 32(\frac{2}{3}\sqrt{\frac{2}{\pi}}\sigma M + M^2/3)$.

Then

$$\mathbb{E}\left[\|f - \hat{f}_n\|^2 \right] \leq \min_{f' \in \Gamma} \left\{ 2\|f - f'\|^2 + \frac{3}{2}\frac{\lambda}{n}\mathrm{pen}(f') \right\} + 3(s+1)\frac{\lambda}{n}. \tag{18}$$

To illustrate the use of Corollary 1 we consider the application of the passive learning scenario to the class of Hölder smooth functions.

Theorem 8. *Let $\Sigma(L, \alpha, M)$ denote the class of functions $f \in \Sigma(L, \alpha)$ that are uniformly bounded, that is $|f(x)| \leq M \ \forall x \in [0,1]^d$. Under the requirements of the passive learning model we have*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{passive}} \sup_{f \in \Sigma(L,\alpha,M)} \mathbb{E}_{f,S_n}[\|\hat{f}_n - f\|^2] \leq C(n/\log n)^{-\frac{2\alpha}{2\alpha+d}} \overset{poly}{\sim} C n^{-\frac{2\alpha}{2\alpha+d}},$$

for n large enough, where $C \equiv C(L, \alpha, M, \sigma^2) > 0$, and $\Theta_{passive}$ denotes the set of all passive estimation strategies.

The proof of Theorem 8 is presented in Appendix D and contains several ingredients used in subsequent results. Note that we obtain the same rate as in the lower bounds of Theorem 1 and Theorem 3 (with respect to the polynomial rate). The logarithmic factor in Theorem 8 is an artifact of the bounding techniques used, and it is due to the fact that we can consider only a countable set of possible estimators in Corollary 1 (see Appendix D for more details). This factor can actually be removed, as well as the bound on the magnitude of f. This requires a much more careful analysis of the error, such as the one done

in [14] for a deterministic sample strategy. In conclusion, the optimal rate of convergence for the class $\Sigma(L, \alpha)$ is the one given in Theorem 3 both for the passive and active learning scenarios.

For the piecewise smooth/constant classes the estimators we consider are built over *Recursive Dyadic Partitions* (RDPs). The elements of an RDP are quasi-disjoint[5] subintervals of $[0,1]^d$, such that their union is the entire unit hypercube. A RDP is any partition that can be constructed using only the following rules:

i) $\{[0,1]^d\}$ is a RDP;

ii) Let $\pi = \{A_0, \ldots, A_{k-1}\}$ be a RDP, where $A_i = [a_{i1}, b_{i1}] \times \ldots \times [a_{id}, b_{id}]$. Then $\pi' = \{A_1, \ldots, A_{i-1}, A_i^{(0)}, \ldots, A_i^{(2^d-1)}, A_{i+1}, \ldots, A_k\}$ is a RDP, where $\{A_i^{(0)}, \ldots, A_i^{(2^d-1)}\}$ is obtained by dividing the hypercube A_i into 2^d quasi-disjoint hypercubes of equal size. Formally, let $q \in \{0, \ldots, 2^d - 1\}$ and $q = q_1 q_2 \ldots q_d$ be the binary representation of q. Then

$$A_i^{(q)} = \left[a_{i1} + \frac{b_{i1} - a_{i1}}{2} q_1, b_{i1} + \frac{a_{i1} - b_{i1}}{2}(1 - q_1)\right] \times \cdots \times \left[a_{id} + \frac{b_{id} - a_{id}}{2} q_d, b_{id} + \frac{a_{id} - b_{id}}{2}(1 - q_d)\right].$$

Whenever a partition π' can be constructed by repeated application of rule (ii) to a partition π we say that the partitions are nested, and that $\pi' \preceq \pi$ (meaning that the partition π' is "finer" than partition π).

Other recursive partition strategies can also be considered, such as "free-split" procedures [3]. These can also be analyzed in our framework, although some extra difficulties arise.

It is clear that an RDP π can be described effectively by a rooted tree structure, where each leaf corresponds to a element of the partition, the root node corresponds to the set $[0,1]^d$, and the internal nodes correspond to the aggregation of elements of π. This idea is illustrated in Figure 1 for the two-dimensional case. Also, a proper RDP, consisting of disjoint sets (instead of quasi-disjoint sets) can be constructed the same way. Denote the set of all RDPs by Π. We define the depth of a leaf in a RDP as the distance (number of links) from the root to the leaf in the tree representation of the RDP. For example in Figure 1(c) the RDP has four leafs at depth two and three leafs at depth one.

The estimates we consider are constructed decorating each of the sets in a RDP with a polynomial of degree $\lfloor \alpha \rfloor$. Formally, let π be a RDP and define

$$\Xi(\pi) = \left\{ g(x) : g(x) = \sum_{A \in \pi} g_A(\boldsymbol{x}) \mathbf{1}_{\boldsymbol{x} \in A},\ g_A \in \mathcal{P}(\lfloor \alpha \rfloor) \right\}. \tag{19}$$

[5]Two sets are quasi-disjoint if and only if their intersection has Lebesgue measure zero.

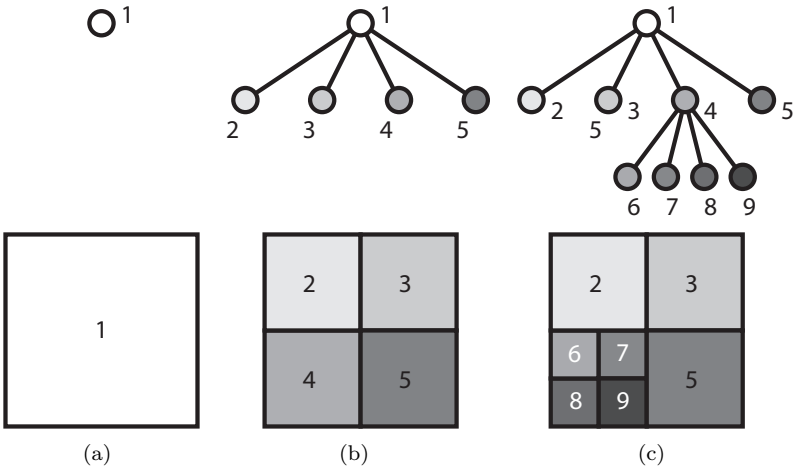


Figure 1: Example of Recursive Dyadic Partitions, and the corresponding tree representations.

where $\mathcal{P}(k)$ is set of all polynomials of degree k on $\mathbb{R}^d$.

The estimator $\hat{f}_n$ we are going to consider is best constructed in a two-stage way. Define $\hat{f}_n^{(\pi)} : [0,1]^d \to \mathbb{R}$ such that

$$\hat{f}_n^{(\pi)} \quad = \quad \arg \min_{g \in \Xi(\pi)} \sum_{i=1}^{n} (Y_i - g(\boldsymbol{X}_i))^2, \tag{20}$$

that is, for a fixed RDP π the function $\hat{f}_n^{(\pi)}$ is the least squares fit of the data to f (over the class $\Xi(\pi)$). Now define

$$\hat{\pi} \equiv \arg \min_{\pi \in \Pi} \left\{ \sum_{i=1}^{n} (Y_i - \hat{f}_n^{(\pi)}(\boldsymbol{X}_i))^2 + \lambda(|\pi|) \right\},$$

where $|\pi|$ denotes the number of elements of partition π, and $\lambda(\cdot)$. Later on we will see that $\lambda(|\pi|) = p(n)|\pi|$. Finally, the estimator $\hat{f}_n : [0,1]^d \to \mathbb{R}$ is defined as

$$\hat{f}_n = \hat{f}_n^{(\hat{\pi})}. \tag{21}$$

The computation of $\hat{f}_n$ can be done by efficiently using tree pruning algorithms, in the spirit of CART [3]. Although the estimator (21) is very appealing and practical, it is difficult to analyze under the scope of Theorem 7 and Corollary 1, since there is an uncountable number of possible estimates (because $\Xi(\pi)$ is uncountable). Instead we are going to analyze a related estimator, where we consider only a finite subset of $\Xi(\pi)$, obtained by quantizing the coefficients of the polynomial decorating the tree leafs. That modified estimator, presented in the Appendix E, allows us to prove the main result of this section.

Theorem 9 (Main Passive Learning Theorem)**.** *Under the passive learning scenario we have*

$$\begin{aligned}\inf_{(\hat{f}_n,S_n)\in\Theta_{passive}}\sup_{f\in PS(L,\alpha,\beta,M)}\mathbb{E}_f[\|\hat{f}_n-f\|^2] \\ \leq \quad C\max\left\{\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}},\left(\frac{n}{\log n}\right)^{-\frac{1}{d}}\right\} \\ \overset{poly}{\sim} \quad C\max\left\{n^{-\frac{2\alpha}{2\alpha+d}},n^{-\frac{1}{d}}\right\}, \end{aligned} \tag{22}$$

where $C\equiv C(L,\alpha,\beta,M,\sigma^2)$.

The proof of the Theorem is presented in Appendix E. We observe that we get the same rate (up to a logarithmic factor) of Proposition 2, therefore this is the optimal rate of convergence for the passive learning scenario.

The above result and estimator are our work-horses in the next section, where we show how these can be used to obtain faster convergence rates in the active learning scenario. Before we proceed we add just one remark

Remark:*Although the estimator used in the proof of Theorem 9 involved a search of all possible RDPs, this is not at all required. We need only to consider RDPs up to a certain depth (more precisely up to depth* $J=\lceil\frac{1}{d}\log(n/\log(n))\rceil$*). This fact is clear from the proof, since in the oracle bound we only need to consider such RDPs.*

To conclude this section we present the results for the piecewise constant class. This is essentially equivalent to Theorem 9 when the smoothness α is taken to infinity.

Theorem 10. *Under the passive learning scenario we have*

$$\begin{aligned}\inf_{(\hat{f}_n,S_n)\in\Theta_{passive}}\sup_{f\in PC(\beta,M)}\mathbb{E}_f[\|\hat{f}_n-f\|^2] \\ \leq \quad C\begin{cases}\frac{\log^2 n}{n} & ,\ if\ \ d=1,\\ \left(\frac{n}{\log n}\right)^{-\frac{1}{d}} & ,\ if\ \ d>1,\end{cases} \\ \overset{poly}{\sim} \quad Cn^{-\frac{1}{d}}, \end{aligned}$$

where $C\equiv C(\beta,M,\sigma^2)$.

A sketch of the proof is presented in Appendix F.

4.2 Active Learning Strategies

In this section we present active learning schemes that improve on the best passive performance rates for the piecewise constant class. The piecewise constant case is somewhat the *canonical case*, in the sense that the regression function f is as simple as possible away from the boundary.

The proposed scheme is based on a two-step approach. In the first step, called the *preview step*, a rough estimator of f is constructed using $n/2$ samples (assume for simplicity that n is even), distributed uniformly over $[0,1]^d$. In the second step, called the *refinement step*, we select $n/2$ samples near the perceived locations of the boundaries (estimated in the preview step) separating constant regions. In the end of this process we will have half the samples concentrated in the vicinity of the boundary set $B(f)$. Since accurately estimating f near the boundary set is key to obtaining faster rates we expect such a strategy to outperform the passive learning technique described earlier. The two-steps of the proposed active learning approach are described in more detail below. For simplicity assume that n is even.

Preview: The goal of this stage is to provide a coarse estimate of the location of $B(f)$. Specifically, collect $n' \equiv n/2$ samples at points distributed uniformly over $[0,1]^d$. Next we proceed by using the passive learning algorithm described in Section 4.1, but we restrict ourselves to RDPs with a leafs at a maximum depth: only RDPs with leafs at a depth $j \leq J = \lceil \frac{d-1}{(d-1)^2+d} \log(n'/\log(n')) \rceil$ are allowed. Denote the set of all such partitions by Π_J. The reason for this choice of depth will be clear from the analysis of the algorithm, but for now notice that these trees are shallower than what is needed to obtain the optimal performance of the passive algorithm (see Appendix F). This creates an estimate whose error is dominated by the squared bias term, and has a small variance. In other words, we obtain a very "stable" coarse estimate of f, where stable means that the estimator does not change much for different realizations of the data. The above strategy ensures that most of the time, leafs that intersect the boundary are at the maximum allowed depth (because otherwise the estimator would incur too much empirical error) and leafs away from the boundary are at shallower depths. Therefore we can "detect" the rough location of the boundary just by looking at the deepest leafs. Denote this estimator by $\hat{f}_0^p$. Formally,

$$\hat{f}_0^p = \hat{f}_{n'}^{(\hat{\pi}_0^p)},$$

where $\hat{f}_{n'}$ was defined in (20) and

$$\hat{\pi}_0^p \equiv \arg\min_{\pi \in \Pi_J} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n'} (Y_i - \hat{f}_{n'}^{(\pi)}(\boldsymbol{X}_i))^2 + \lambda(|\pi|) \right\}.$$

Unfortunately, if the set $B(f)$ is somewhat aligned with the dyadic splits of the RDP, leafs intersecting the boundary can be pruned without incurring a large error, and therefore we wouldn't be able to properly detect the boundary in those situations. This is illustrated in Figure 2a; the highlighted cell was pruned and contains a piece of the boundary. The error incurred by pruning should be small, since that region is mostly of a constant region. However,worst-case analysis reveals that the squared bias induced by these small volumes can add up, precluding the desired rates.

The cause of this problem is the fact that our preview estimator is not translation invariant; that is, if we consider in the observation model (A1) a

spatially translated version of f, then our RDP-based estimate may changes considerably. A way of mitigating this issue is to consider multiple RDP-based estimators, each one using a RDP appropriately shifted. We use $d+1$ estimators in the preview step: one on the initial uniform partition, and d over partitions whose dyadic splits have been translated by 2^{-J} in each one of the d coordinates. The main idea is illustrated in Figure 2: For $\hat{f}_1^p$ pruning the cells intersecting the highlighted boundary region would cause a large error, therefore making it easier to detect the boundary.

Formalizing the structure of the shifted partitions is a little cumbersome, but for the sake of completeness we include the rules to construct such partitions below. These are similar the rules presented before for the regular RDPs. A shifted RDP in the l^{th} coordinate satisfies the following.

i) $\{[0,1]^d\}$ is a RDP;

ii) Let $\pi = \{A_1, \ldots, A_k\}$ be a RDP, where $A_i = [a_{i1}, b_{i1}] \times \ldots \times [a_{id}, b_{id}]$. Then $\pi' = \{A_1, \ldots, A_{i-1}, A_i^{(1)}, \ldots, A_i^{(2^d)}, A_{i+1}, \ldots, A_k\}$ is a RDP, where $\{A_i^{(1)}, \ldots, A_i^{(2^d)}\}$ is obtained by dividing the hypercube A_i into 2^d quasi-disjoint hypercubes of equal size (expect near the edge of the unit hypercube). Formally, let $q \in \{0, \ldots, 2^{d-1}\}$ and $q = q_1 q_2 \ldots q_d$ be the corresponding binary representation. Then

$$
\begin{aligned}
A_i^{(q)} = &\left[a_{i1} + \frac{b_{i1} - a_{i1}}{2} q_1, b_{i1} + \frac{a_{i1} - b_{i1}}{2}(1 - q_1)\right] \times \cdots \\
&\times \left[a_{il} + \frac{b_{il} - a_{il}}{2} q_l + 2^{-J-1}(\mathbf{1}_{a_{il}=0} + \mathbf{1}_{b_{il}=1}) q_l, \right. \\
&\quad \left. b_{il} + \frac{a_{il} - b_{il}}{2}(1 - q_l) + 2^{-J-1}(\mathbf{1}_{a_{il}=0} + \mathbf{1}_{b_{il}=1})(1 - q_l)\right] \times \cdots \\
&\left[a_{id} + \frac{b_{id} - a_{id}}{2} q_d, b_{id} + \frac{a_{id} - b_{id}}{2}(1 - q_d)\right].
\end{aligned}
$$

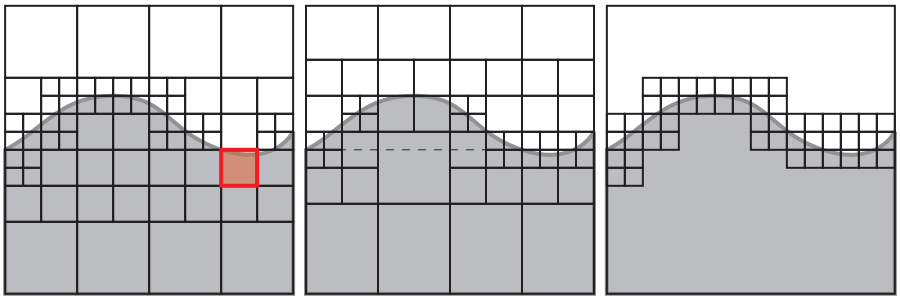
Denote the set of all l-coordinate shifted RDPs with all leafs at depths no greater than J by $\Pi_J^{l_{\text{shift}}}$. The preview estimators built on shifted partitions are defined as

$$\hat{f}_l^p \equiv \hat{f}_{n'}^{(\hat{\pi}_l^p)},$$

where

$$\hat{\pi}_l^p \equiv \arg \min_{\pi \in \Pi_J^{l_{\text{shift}}}} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n'} (Y_i - \hat{f}_{n'}^{(\pi)}(\boldsymbol{X}_i))^2 + \lambda(|\pi|) \right\}.$$

The analysis of an estimator built on top of this shifted partitions is similar to the one for the regular partitions. Therefore the proof of Theorem 10 applies also to this estimator. The only difference is that now the volume of cells in the partition at depth j might be larger than 2^{-j}, although it is at most 2×2^{-j} (therefore the right-hand-side of (43) is multiplied by 2). This only affects the constant C in Theorem 10.



(a) (b) (c)

Figure 2: Illustration of the shifted RDP construction for $d = 2$: (a) RDP used in $\hat{f}_0^p$. The highlighted cell intersects the boundary but it was pruned, since the pruning does not incur in severe error. (b) Shifted RDP, used in $\hat{f}_1^p$. In this case there is no pruning, since it would cause a large error. (c) These are the cells that are going to be refined in the refinement stage.

Any leaf that is at the maximum depth of any of the $d+1$ RDPs pruned in the preview step indicates the highly probable presence of a boundary, and will be refined in the next stage.

Refinement: With high probability, the boundary is contained in leafs at the maximum depth. In the refinement step we collect additional n' samples in the corresponding partition sets, and obtain a refined estimate of the function f. Let

$$\mathcal{R} = \bigcup_{l=0}^{d} \{A \in \hat{\pi}_l^p : \ A \text{ corresponds to a leaf at depth } J\}. \tag{23}$$

Note that there might be repetitions in the elements of $\mathcal{R}$ with the above definition. In the following we assume that those repetitions are removed. For each set $A \in \mathcal{R}$ we collect $n'/|\mathcal{R}|$ samples, distributed uniformly over each A. Therefore we collect a total of n' samples in this step. For each set $A \in \mathcal{R}$ we repeat the tree pruning process described in Section 4.1 (but now instead of defining the possible RDPs over the unit hypercube, we define them over A). This produces a higher resolution estimate in the vicinity of the boundary set $B(f)$, yielding a better performance than the passive learning technique. Denote the estimators obtained over the set $A \in \mathcal{R}$ by $\hat{f}_A^r$. The overall estimator of f, obtained after the preview and refinement steps is denoted by $\hat{f}_{\text{active}}$ and it is defined as

$$\hat{f}_{\text{active}}(\boldsymbol{x}) = \begin{cases} \hat{f}_A^r(\boldsymbol{x}) & \text{, if } \boldsymbol{x} \in A : A \in \mathcal{R}, \\ \hat{f}_0^p(\boldsymbol{x}) & \text{, otherwise} \end{cases}.$$

To guarantee the performance of this algorithm we need a further assumption on the regression function f, namely:

(A4) Let $f \in PC(\beta, M)$ and consider the partition of $[0,1]^d$ into m^d identical hypercubes, with $m = 2^J$ and J as defined above. Let $f_J : [0,1]^d \to \mathbb{R}$ be a coarse approximation of f. Formally, for all partition sets $A \in \pi_J$ we have

$$f_J(\boldsymbol{x}) = m^d \sum_{A \in \pi_J} \left(\int_A f(\boldsymbol{t}) \mathrm{d}\boldsymbol{t} \right) \mathbf{1}_A(\boldsymbol{x}).$$

Let $A \in \pi_J$ such that $A \cap B(f) \neq \emptyset$. Consider now the l-coordinate shifted RDPs, and let $\mathcal{A}(A, l)$ denote the cell at depth $J-1$ containing A (recall that $l = 0$ corresponds to the usual non-shifted RDPs). We require that, for at least one $l \in \{0, \ldots, d\}$,

$$\int_{\mathcal{A}(A,l)} \left(f_J(\boldsymbol{x}) - \frac{1}{|\mathcal{A}(A,l)|} \int_{\mathcal{A}(A,l)} f(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \right)^2 \mathrm{d}\boldsymbol{x} \geq C_b(f) 2^{-dJ}, \qquad (24)$$

where $C_b(f)$ and $n \geq n_0(f)$, with $n_0(f) > 0$.

Although this condition restricts the shape of the boundary sets, it is still quite general, and encompasses many interesting cases, in particular Lipshitz boundary fragments. Essentially (A4) require $B(f)$ to be "cusp-free"[6]. A cusp-like boundary is difficult to detect with our preview step, since it is very "thin", but might still have enough volume to prevent the algorithm from achieving the correct rate. We conjecture that it is possible to attain the active minimax rates without requiring (A4), but a different algorithm may be needed. Figure 3 illustrates what happens when a boundary is present. The piece of the boundary in the central cell is "sensed" by the three different RDPs. Clearly the "green" partition (with an arrow) is able to detect the boundary piece in the center cell.

Theorem 11. *Under the active learning scenario we have, for $d \geq 2$ and functions f satisfying (A4),*

$$\mathbb{E}\left[\|\hat{f}_{active} - f\|^2 \right] \leq C \left(\frac{n}{\log n} \right)^{-\frac{1}{d-1+1/d}} \overset{poly}{\sim} C n^{-\frac{1}{d-1+1/d}},$$

where $C > 0$.

The proof of Theorem 11 is presented in Appendix G. Note that we improve on the passive rates using this technique, but do not achieve the lower bound of Theorem 5. By reiterating this approach (generalizing to a multi-step approach) it is possible get arbitrarily close to the rate of Theorem 5, as we see below. This bound does not hold uniformly over the entire class of functions, unlike all of our previews results. This is again a limitation of our algorithm, and the cause is the *ribbon* problem. If f is a thin ribbon then, unless there is enough resolution

[6] A cusp-free boundary cannot have the behavior you observe in the graph of $\sqrt{|x|}$ at the origin. Less "aggressive" kinks are allowed, such as in the graph of $|x|$.

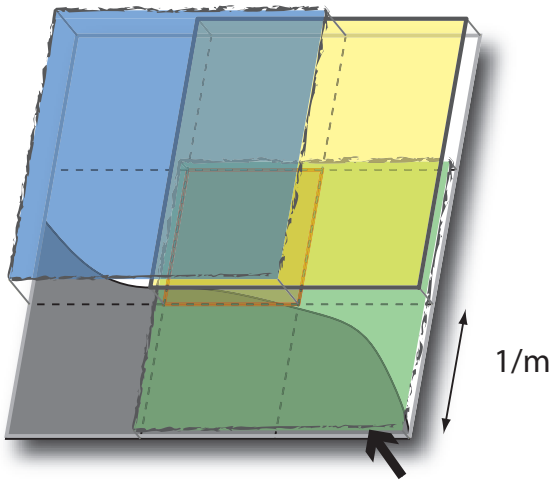


Figure 3: Illustration of condition (24). The "green" RDP cell (with arrow) is able to "feel" the shaded boundary in the central cell.

available (n is large enough), we cannot detect that ribbon in the preview step. Therefore, for n below some critical value the performance of the algorithm is actually worst than the passive learning algorithm described in Section 4.1. For n above this critical value we are able to detect the boundary set, and therefore the active strategy of the algorithm becomes beneficial.

As said before, one can reiterate the two-step procedure: For example, to obtain a three step procedure we can start with a similar preview step, using $n'' \equiv n/3$ samples, and a different value of J, adjusted accordingly. In the refinement step apply the two-step procedure as described above, instead of the passive strategy. With this three step approach we attain the error decay rate

$$\overset{\text{poly}}{\sim} n^{\frac{1}{d-1+\frac{1}{(d-1)^2+d}}},$$

where in the first step we used

$$J = \lceil \frac{(d-1)^2}{d(d^2-2d+2)} \log(n''/\log(n'')) \rceil.$$

Notice that the error decay rate improved with respect to Theorem 11. This procedure can be repeated, and we get the following result

Theorem 12. *Using the reiterated scheme described above we have, for $d \geq 2$ and functions f satisfying (A4),*

$$\mathbb{E}\left[\|\hat{f}_{active} - f\|^2\right] \overset{poly}{\leq} Cn^{-\frac{1}{d-1+\epsilon}},$$

where $\epsilon > 0$ depends on the total number of steps of the algorithm, and can be made arbitrarily small as the number of steps increases.

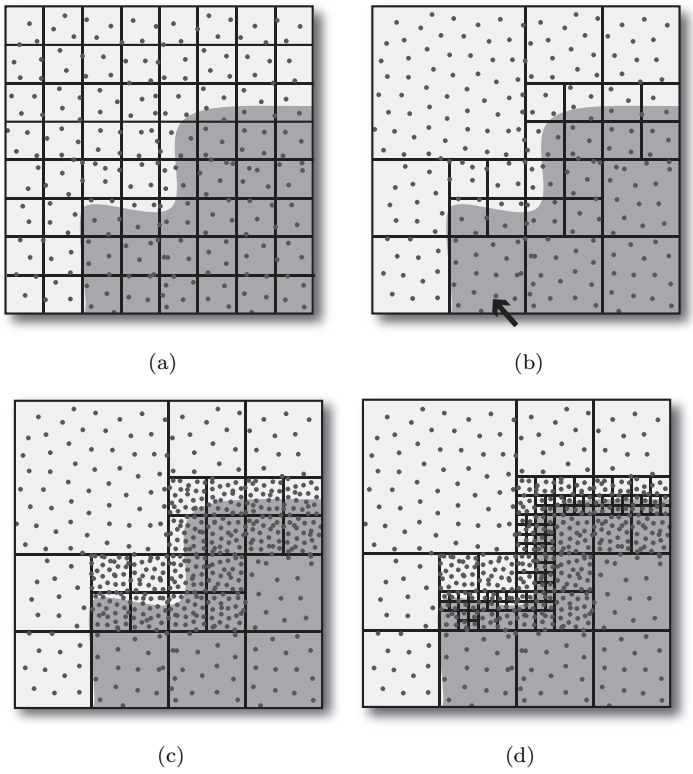


Figure 4: The two step procedure for $d = 2$ (no shifted partitions): (a) Initial unpruned RDP and $n/2$ samples. (b) Preview step RDP. Note that the cell with the arrow was pruned, but it contains a part of the boundary. (c) Additional sampling for the refinement step. (d) Refinement step.

The proof of Theorem 12 goes by careful choice of the maximum resolution J at each step, by balancing the error between each subsequent steps (as in Theorem 11).

5 Final Remarks and Open Questions

The results presented in this paper show that in certain scenarios active learning attains provable gains over the classical passive approaches. Active learning is an intuitively appealing idea and may find application in many practical problems. Despite these draws, the analysis of such active methods is quite challenging due to the loss of statistical independence in the observations (recall that now the sample locations are coupled with all the observations made in the past). The function classes presented here are non-trivial canonical examples illustrating under what conditions one might expect active learning to improve rates of

convergence. The algorithm presented here for actively learning members of the piecewise constant class demonstrates the possibilities of active learning. In fact, this algorithm has already been applied successfully in the context of field estimation using wireless sensor networks [20].

The algorithmic ideas presented were in this paper might be simple and intuitive, but show the difficulties inherent a sound analysis of such methods. For example, the algorithm developed is rather "aggressive": In the preview step we cannot miss any part of the boundary set, since we have no further chance of detecting it (in the refinement step). This is the reason we need assumption (A4). Although less aggressive algorithmic techniques can be devised, their analysis becomes extremely difficult. The algorithm proposed can also be extended to the piecewise smooth class of functions, as long as condition (24) is satisfied. This enforces that such a function f is discontinuous in the boundary set, and so our techniques still work.

Another area that can greatly benefic from the ideas behind active learning and boundary estimation is binary classification: The goal in this context is to learn the Bayes decision boundary, so clearly only certain parts of the feature space are relevant for this task. There are various pieces of work presenting ideas and methods for active learning in this scenario, but a solid theoretical framework is still largely undeveloped.

Acknowledgements

The authors would like to thank Jarvis Haupt for the helpful discussions regarding the Bernstein moment condition, and Aarti Singh for pointing out numerous incorrect statements in the proofs.

A Proof of Theorem 3

The proof of Theorem 3 is closely mimics the proof of Theorem 1 presented in [19] in the 1-dimensional setting. The key idea of the proof is to reduce the problem of estimating a function in $\Sigma(L, \alpha)$ to the problem of deciding among a finite number of hypothesis. The proof methodology for the passive setting works for the active scenario because we can choose an adequate set of hypothesis without knowledge of the sampling strategy. There is also another modification needed, due to the extra flexibility of the sampling strategy. For the sake of completeness we include the entire proof in this paper.

The proof is essentially the application of the following theorem.

Theorem 13 (Main Theorem of Risk Minimization (Kullback divergence version))**.** *Let Θ be a class of models. Associated with each model $\theta \in \Theta$ we have a probability measure P_θ. Let $M \geq 2$ be an integer and let $d(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R}$ be a semidistance. Suppose we have $\{\theta_0, \ldots, \theta_M\} \in \Theta$ such that*

i) $d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall_{0 \leq j,k \leq M}$,

ii) $P_{\theta_j} \ll P_{\theta_0}, \quad \forall_{j=1,\dots,M}$,

iii) $\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(P_{\theta_j} \| P_{\theta_0}) \le \gamma \log M$,

where $0 < \gamma < 1/8$. *The following bound holds.*

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} P_\theta \left(d(\hat{\theta}, \theta) \ge s \right) \ge \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\gamma - 2\sqrt{\frac{\gamma}{\log M}} \right) > 0,$$

where the infimum is taken with respect to the collection of all possible estimators of θ, *and* KL *denotes the Kullback-Leibler divergence* [7].

From the theorem with can immediately show the following corollary.

Corollary 2. *Under the assumptions of Theorem 13 we have*

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}[d^2(\hat{\theta}, \theta)] \ge s^2 \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\gamma - 2\sqrt{\frac{\gamma}{\log M}} \right) > cs^2,$$

for some $c(\gamma, M) > 0$.

Proof. The result follows from straightforward application of Markov's inequality,

$$P_\theta \left(d(\hat{\theta}, \theta) \ge s \right) = P_\theta \left(d^2(\hat{\theta}, \theta) \ge s^2 \right) \le \frac{1}{s^2} \mathbb{E}[d^2(\hat{\theta}, \theta)].$$

□

Although in our formulation of the problem we assume (A2) that the additive noise is Gaussian, for the proof of the minimax lower bounds we can use a more general assumption, namely we require that there exists $p^* > 0$ and $v_0 > 0$ such that

$$\mathrm{KL}(p_W(\cdot) \| p_W(\cdot + v)) = \int \log \frac{p_W(w)}{p_W(w+v)} p_W(w) \mathrm{d}\nu(w) \le p^* v^2, \quad \forall\ |v| \le v_0.$$

For the Gaussian model this condition is satisfied with $p^* = 1/(2\sigma^2)$ and $v_0 = \infty$.

Consider a fixed sample size n. To apply Theorem 13 to the problem considered we need to construct a collection of hypothesis $f_j(\cdot) \in \Sigma(L, \alpha)$, $j = 0, \dots, M$. Let $c_0 > 0$ and define

$$m = \left\lceil c_o n^{\frac{1}{2\alpha+d}} \right\rceil, \text{[8]} \quad h = \frac{1}{m}, \quad \boldsymbol{x}_{\boldsymbol{k}} = \frac{\boldsymbol{k} - 1/2}{m},$$

[7] Let P and Q be two probability measures defined on a probability space $(\mathcal{X}, \mathcal{B})$. The Kullback-Leibler divergence is defined as

$$\mathrm{KL}(P\|Q) = \begin{cases} \int \log \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P & \text{, if } P \ll Q, \\ +\infty & \text{, otherwise.} \end{cases},$$

where $\mathrm{d}P/\mathrm{d}Q$ is the Radon-Nikodym derivative of measure P with respect to measure Q.

[8] $k = \lceil x \rceil$ is the minimal integer such that $x < k$.

and

$$\varphi_{\boldsymbol{k}}(\boldsymbol{x}) = Lh^{\alpha} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_{\boldsymbol{k}}}{h}\right),$$

where $\boldsymbol{k} \in \{1,\ldots,m\}^d$, $\boldsymbol{x} \in [0,1]^d$ and $K : \mathbb{R}^d \to [0,+\infty)$ satisfies $K \in \Sigma(1,\alpha)$ and supp $K = (-1/2,1/2)^d$. It is easily shown that such a function K exists [9]. Let $\Omega = \{\boldsymbol{\omega} = (\omega_1,\ldots,\omega_{m^d}),\ \omega_i \in \{0,1\}\} = \{0,1\}^{m^d}$, and define

$$\xi = \{f_{\boldsymbol{\omega}}(\cdot) : f_{\boldsymbol{\omega}}(\cdot) = \sum_{\boldsymbol{k}\in\{1,\ldots,m\}^d} \omega_{\boldsymbol{k}}\varphi_{\boldsymbol{k}}(\cdot),\ \boldsymbol{\omega} \in \Omega\}.$$

Note that $\varphi_{\boldsymbol{k}} \in \Sigma(L,\alpha)$ and these functions have disjoint support, therefore $\xi \subseteq \Sigma(L,\alpha)$.

For $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$

$$\begin{aligned}
d(f_{\boldsymbol{\omega}}, f'_{\boldsymbol{\omega}}) &= \left[\int_{[0,1]^d} (f_{\boldsymbol{\omega}}(\boldsymbol{x}) - f'_{\boldsymbol{\omega}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}\right]^{1/2} \\
&= \left[\sum_{\boldsymbol{k}\in\{1,\ldots,m\}^d} (\omega_{\boldsymbol{k}} - \omega'_{\boldsymbol{k}})^2 \int_{[0,1]^d} \varphi^2_{\boldsymbol{k}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right]^{1/2} \\
&= Lh^{\alpha+d/2}\|K\| \left[\sum_{\boldsymbol{k}\in\{1,\ldots,m\}^d} |\omega_{\boldsymbol{k}} - \omega'_{\boldsymbol{k}}|\right]^{1/2} \\
&= Lh^{\alpha+d/2}\|K\| \sqrt{\rho(\boldsymbol{\omega},\boldsymbol{\omega'})},
\end{aligned}$$

where ρ is the Hamming distance between $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$, and $\|K\| = \sqrt{\int_{[0,1]^d} K^2(\boldsymbol{x})\mathrm{d}x}$. We will choose our hypotheses from ξ, but we do not need the entire set. We need the following result from information theory.

Lemma 1 (Varshamov-Gilbert bound, 1962)**.** *Let $m^d \geq 8$. There exists a subset $\{\boldsymbol{\omega}^{(0)}, \boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)}\}$ of Ω such that $\boldsymbol{\omega}^{(0)} = (0,\ldots,0)$ and*

$$\rho(\boldsymbol{\omega}^{(j)}, \boldsymbol{\omega}^{(k)}) \geq m^d/8, \quad \forall\ 0 \leq j < k \leq M$$

and $M \geq 2^{m^d/8}$.

For a proof of the Lemma 1 see [19].

Finally let $\{\boldsymbol{\omega}^{(0)}, \boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)}\}$ be a set satisfying the conditions of Lemma 1 and define $f_j(\cdot) \equiv f_{\boldsymbol{\omega}^{(j)}}(\cdot)$, with $j = 0,\ldots,M$. This is the collection of hypotheses we will use with Theorem 13. We need to verify the various conditions in the theorem. As already pointed out, notice that $f_j \in \Sigma(L,\alpha)$.

[9] For example

$$K(\boldsymbol{x}) = a\tilde{K}(2\boldsymbol{x}), \text{ with } \quad \tilde{K}(\boldsymbol{x}) = \prod_{i=1}^{d} \exp(-\frac{1}{1-x_i^2})\mathbf{1}(|x_i| < 1).$$

where $\boldsymbol{x} = (x_1,\ldots,x_i)$ and $a > 0$ is sufficiently small.

i)

$$\begin{aligned} d(f_j, f_k) = Lh^{\alpha+d/2}\|K\|\sqrt{\rho(\boldsymbol{\omega^{(j)}}, \boldsymbol{\omega^{(k)}})} &\geq Lh^{\alpha+d/2}\|K\|\sqrt{m^d/8} \\ &= Lm^{-\alpha}\|K\|/\sqrt{8}, \end{aligned}$$

as long as $m^d > 8$, as a result of Lemma 1. This is the case if $n \geq n^*$ with $n^* = (8^{1/d}/c_0)^{2\alpha+d}$.

Let $n \geq n^*$, then $m \leq c_0 n^{\frac{1}{2\alpha+d}}(1 + 8^{-1/d})$, and therefore

$$\begin{aligned} d(f_j, f_k) &\geq L\frac{1}{\sqrt{8}}m^{-\alpha}\|K\| \\ &\geq L\frac{1}{\sqrt{8}}(1 + 8^{-1/d})^{-\alpha}c_0^{-\alpha}\|K\|n^{-\frac{\alpha}{2\alpha+d}} \\ &\geq L/3\ c_0^{-\alpha}\|K\|n^{-\frac{\alpha}{2\alpha+d}} \\ &= A\psi_n, \end{aligned}$$

where $\psi_n = n^{-\frac{\alpha}{2\alpha+d}}$ and $A = Lc_0^{-\alpha}\|K\|/3$.

ii) Under the active sampling modeling assumption we see that the probability measure of $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}, Y_1, \ldots, Y_n)$ has a nice factorization. To avoid cumbersome technical derivations assume that the conditional random variable $\boldsymbol{X_i}|\boldsymbol{X_1}\ldots\boldsymbol{X_{i-1}}, Y_1\ldots Y_{i-1}$ has a density $p_{\boldsymbol{X_i}|\boldsymbol{X_1}\ldots\boldsymbol{X_{i-1}},Y_1\ldots Y_{i-1}}$ with respect to a suitable dominating measure. For a function $f \in \Sigma(L, \alpha)$ the joint probability measure of the sample points and observations has a density (with respect to a suitable dominating measure) of the form

$$\begin{aligned} &p_{\boldsymbol{X_1},\ldots,\boldsymbol{X_n},Y_1,\ldots,Y_n}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}, y_1, \ldots, y_n) \\ &\quad = p_{\boldsymbol{Z}_n^X,\boldsymbol{Z}_n^Y}(\boldsymbol{z}_n^X, \boldsymbol{z}_n^Y) \\ &\quad = \prod_{i=1}^{n} p_{Y_i|\boldsymbol{X_i};f}(y_i|\boldsymbol{x_i})\ p_{\boldsymbol{X_i}|\boldsymbol{Z}_{i-1}^X,\boldsymbol{Z}_{i-1}^Y}(\boldsymbol{x_i}|\boldsymbol{z}_{i-1}^X, \boldsymbol{z}_{i-1}^Y), \end{aligned} \tag{25}$$

where $\boldsymbol{Z}_i^X \equiv (\boldsymbol{X_1}, \ldots, \boldsymbol{X_i})$, $\boldsymbol{z}_i^X \equiv (\boldsymbol{x_1}, \ldots, \boldsymbol{x_i})$, $\boldsymbol{Z}_i^Y \equiv (Y_1, \ldots, Y_i)$ and $\boldsymbol{z}_i^Y \equiv (y_1, \ldots, y_i)$. Now taking into account (A2) we have $p_{Y_i|\boldsymbol{X_i};f}(y_i|\boldsymbol{x_i}) = p_W(y_i - f(\boldsymbol{x_i}))$ and we are done since all these measures are absolutely continuous with respect to each other.

iii) Note that, for n large enough, $f_j(\cdot) \leq Lh^\alpha K_{\max} \leq v_0$, where $K_{\max} = \max_{\boldsymbol{x}\in\mathbb{R}^d} K(\boldsymbol{x})$. Namely the above holds if $n > n_0$, with $n_0 = c_0^{-(2\alpha+d)}(LK_{\max}/v_0)^{(2\alpha+d)/\alpha}$. Taking into account the factorization in

(25) and recalling that $f_0(\cdot) = 0$ we have, for $n > n_0$,

$$
\begin{aligned}
&\mathrm{KL}(P_j \| P_0) \\
&= \mathbb{E}_{f_0}\left[\log \frac{\prod_{i=1}^n p_{Y_i|\boldsymbol{X_i};f_j}(Y_i|\boldsymbol{X_i})\; p_{\boldsymbol{X_i}|\boldsymbol{Z}_{i-1}^X,\boldsymbol{Z}_{i-1}^Y}(\boldsymbol{X_i}|\boldsymbol{Z}_{i-1}^X,\boldsymbol{Z}_{i-1}^Y)}{\prod_{i=1}^n p_{Y_i|\boldsymbol{X_i};f_0}(Y_i|\boldsymbol{X_i})\; p_{\boldsymbol{X_i}|\boldsymbol{Z}_{i-1}^X,\boldsymbol{Z}_{i-1}^Y}(\boldsymbol{X_i}|\boldsymbol{Z}_{i-1}^X,\boldsymbol{Z}_{i-1}^Y)}\right] \\
&= \mathbb{E}_{f_0}\left[\log \frac{\prod_{i=1}^n p_{Y_i|\boldsymbol{X_i};f_j}(Y_i|\boldsymbol{X_i})}{\prod_{i=1}^n p_{Y_i|\boldsymbol{X_i};f_0}(Y_i|\boldsymbol{X_i})}\right] \\
&= \mathbb{E}_{f_0}\left[\sum_{i=1}^n \log \frac{p_{Y_i|\boldsymbol{X_i};f_j}(Y_i|\boldsymbol{X_i})}{p_{Y_i|\boldsymbol{X_i};f_0}(Y_i|\boldsymbol{X_i})}\right] \\
&= \sum_{i=1}^n \mathbb{E}_{f_0}\left[\log \frac{p_W(Y_i - f_j(\boldsymbol{X_i}))}{p_W(Y_i - f_0(\boldsymbol{X_i}))}\right] \\
&= \sum_{i=1}^n \mathbb{E}_{f_0}\left[\mathbb{E}_{f_0}\left[\log \frac{p_W(Y_i - f_j(\boldsymbol{X_i}))}{p_W(Y_i - f_0(\boldsymbol{X_i}))}\middle| \boldsymbol{X_1},\ldots,\boldsymbol{X_n}\right]\right] \\
&\leq \sum_{i=1}^n \mathbb{E}_{f_0}\left[p^*(f_j(\boldsymbol{X_i}) - f_0(\boldsymbol{X_i}))^2\right] \\
&= \sum_{i=1}^n \mathbb{E}_{f_0}\left[p^* f_j^2(\boldsymbol{X_i})\right] \\
&\leq \sum_{i=1}^n p^* L^2 h^{2\alpha} K_{\max}^2 \\
&= p^* L^2 K_{\max}^2 n m^{-2\alpha} \\
&\leq p^* L^2 K_{\max}^2 n c_0^{-2\alpha} n^{-\frac{2\alpha}{2\alpha+d}} \\
&= p^* L^2 K_{\max}^2 c_0^{-(2\alpha+d)} c_0^d n^{\frac{d}{2\alpha+d}} \leq p^* L^2 K_{\max}^2 c_0^{-(2\alpha+d)} m^d.
\end{aligned}
$$

From Lemma 1 we have $m^d \leq 8 \log M / \log 2$ therefore choosing

$$c_0 = \left(\frac{8 p^+ L^2 K_{\max}^2}{\gamma \log 2}\right)^{\frac{1}{2\alpha+d}},$$

with $0 < \gamma < 1/8$ yields the desired result.

Since all the conditions of Theorem 13 are met we can apply Corollary 2 and conclude that, for $n > \max\{n^*, n_0\}$,

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{\text{active}}} \sup_{f \in \Sigma(L,\alpha)} \mathbb{E}_{f,S_n}[d^2(\hat{f}_n, f)] \geq A^2 n^{-\frac{2\alpha}{2\alpha+d}}, \tag{26}$$

where $A^2 = (L/3)^2 \|K\|^2 \left(\frac{8 p^+ L^2 K_{\max}^2}{\gamma \log 2}\right)^{\frac{-2\alpha}{2\alpha+d}} < (L/3)^2 \|K\|^2 \left(\frac{64 p^+ L^2 K_{\max}^2}{\log 2}\right)^{\frac{-2\alpha}{2\alpha+d}}$. □

B Proof of Theorem 7

The proof follows closely the strategy in [1], with minor changes pertaining the different noise model considered. For the sake of completeness we include the full derivation here.

The proof hinges on a concentration inequality due to Craig [7].

Theorem 14 (Craig, 1933)**.** *Let $\{U_i\}_{i=1}^n$ be independent random variables, satisfying the Bernstein moment condition*

$$\mathbb{E}\left[|U_i - E[U_i]|^k\right] = Var(U_i)\frac{k!}{2}h^{k-2},$$

for some $h > 0$ and all $k \geq 2$. Let $\overline{U} = (1/n)\sum_{i=1}^n U_i$. Then

$$\Pr\left(\overline{U} - \mathbb{E}[\overline{U}] \geq \frac{\tau}{n\epsilon} + \frac{n\epsilon\, Var(\overline{U})}{2(1-c)}\right) \leq \exp(-\tau),$$

for $0 < \epsilon h \leq c < 1$ and $\tau > 0$.

Start by defining

$$r(f', f) = \mathbb{E}\left[(Y - f'(\boldsymbol{X}))^2\right] - \mathbb{E}\left[(Y - f(\boldsymbol{X}))^2\right].$$

Note that

$$r(f', f) = \mathbb{E}\left[(f'(\boldsymbol{X}) - f(\boldsymbol{X}))^2\right],$$

since $\mathbb{E}[Y|\boldsymbol{X}] = f(\boldsymbol{X})$. Define now the empirical version of $r(f', f)$, that is

$$\begin{aligned}\hat{r}_n(f', f) &\equiv \frac{1}{n}\sum_{i=1}^n (Y_i - f'(\boldsymbol{X}_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f(\boldsymbol{X}_i))^2 \\ &= -\frac{1}{n}\sum_{i=1}^n U_i,\end{aligned}$$

where $U_i = -(Y_i - f'(\boldsymbol{X}_i))^2 + (Y_i - f(\boldsymbol{X}_i))^2$. Notice that the estimator in (15) can be written as

$$\hat{f}_n(\boldsymbol{X}, \boldsymbol{Y}) = \arg\min_{f' \in \Gamma}\left\{\hat{r}(f', f^*) + \frac{\lambda}{n}\mathrm{pen}(f')\right\}.$$

At this point we are going to apply Theorem 14 to $\{U_i\}_{i=1}^n$. For this we need to verify the moment condition in the Theorem. Begin by noticing that

$$\begin{aligned}U_i &= 2(Y_i - f(\boldsymbol{X}_i))(f'(\boldsymbol{X}_i) - f(\boldsymbol{X}_i)) - (f(\boldsymbol{X}_i) - f'(\boldsymbol{X}_i))^2 \qquad (27)\\ &= 2W_i(f'(\boldsymbol{X}_i) - f(\boldsymbol{X}_i)) - (f(\boldsymbol{X}_i) - f'(\boldsymbol{X}_i))^2 \qquad (28)\end{aligned}$$

The variance of U_i can be easily upper bounded noticing that the U_i is the sum of two uncorrelated terms. The variance of the first term is

$$\begin{aligned}\mathrm{Var}\left(2W_i\left(f'(\boldsymbol{X}_i) - f(\boldsymbol{X}_i)\right)\right) &= 4\mathrm{Var}(W_i)\mathbb{E}\left[(f'(\boldsymbol{X}_i) - f(\boldsymbol{X}_i))^2\right] \\ &= 4\mathrm{Var}(W_i)r(f', f).\end{aligned}$$

The variance of the second term is easily bounded by

$$\begin{aligned}\mathrm{Var}\left((f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2\right) &\leq \mathbb{E}\left[(f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^4\right]\\ &\leq 4M^2\mathbb{E}\left[(f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2\right]\\ &\leq 4M^2 r(f',f),\end{aligned}$$

therefore we conclude that $\mathrm{Var}(U_i) \leq 4(\sigma^2+M^2)r(f',f)$.

To determine the moment condition constant h we will use a result presented in [11]. Let A and B be two uncorrelated random variables satisfying the moment condition with constants h_A and h_B respectively. Then $A+B$ satisfies the moment condition with constant $8(h_A+h_B)$. We now proceed by splitting U_i in two terms, as in the computation of the variance, and checking the moment condition for each one of the these. For the first term we have

$$\begin{aligned}&\mathbb{E}\left[\left|2W_i(f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))\right|^k\right]\\ &= \mathbb{E}\left[|2W_i|^k\right]\mathbb{E}\left[|f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i)|^k\right]\\ &\leq \mathrm{Var}(2W_i)\frac{k!}{2}(2h)^{k-2}\mathbb{E}\left[|f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i)|^2\right](2M)^{k-2}\\ &\leq \mathrm{Var}\left(2W_i\left(f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i)\right)\right)\frac{k!}{2}(4hM)^{k-2},\end{aligned}$$

for $k \geq 2$. The second term is bounded, and so we have simply

$$\begin{aligned}&\mathbb{E}\left[\left|(f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2-\mathbb{E}\left[(f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2\right]\right|^k\right]\\ &\leq \mathrm{Var}\left((f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2\right)(4M^2)^{k-2}\\ &\leq \mathrm{Var}\left((f'(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2\right)\frac{k!}{2}(4M^2/3)^{k-2},\end{aligned}$$

for $k \geq 2$. Finally, using the result in [11] we conclude that U_i satisfies Bernstein's moment condition with

$$h_{U_i} = 8(4hM+4M^2/3).$$

Applying Theorem 14 to $\{U_i\}_{i=1}^n$, with $\tau = \mathrm{pen}(f') + \log(1\delta)$ and $\epsilon = 1/\lambda$ we get

$$r(f',f) - \hat{r}_n(f',f) \geq \lambda\frac{\mathrm{pen}(f')+\log(1/\delta)}{n} + \frac{2(\sigma^2+M^2)r(f',f)}{\lambda(1-c)},$$

with probability not greater than $\delta e^{-\mathrm{pen}(f')}$. Using the union of events bound we conclude that

$$r(f',f) - \hat{r}_n(f',f) < \lambda\frac{\mathrm{pen}(f')+\log(1/\delta)}{n} + \frac{2(\sigma^2+M^2)r(f',f)}{\lambda(1-c)}, \qquad (29)$$

for all $f' \in \Gamma$, with probability at least $1-s\delta$. We need to choose c and λ so that the conditions in Theorem 14 hold, therefore take $c = \epsilon h_{U_i} = 32(hM+M^2/3)/\lambda$

and $\lambda > h_{U_i}$ (so that $c < 1$). Rearranging the terms in (29) we get

$$(1-a)r(f',f) < \hat{r}_n(f',f) + \frac{\lambda}{n}\mathrm{pen}(f') + \frac{\lambda}{n}\log(1/\delta), \tag{30}$$

with probability at least $1 - s\delta$, where $a = \frac{2(\sigma^2+M^2)}{\lambda(1-c)}$. For out purposes it is desirable that $a < 1$. This can be ensured by taking $\lambda > 2(\sigma^2+M^2)+32(hM+M^2/3)$.

Taking into account the definition of $\hat{f}_n$ we have in particular that

$$\begin{aligned}(1-a)r(\hat{f}_n,f) &< \hat{r}_n(\hat{f}_n,f) + \frac{\lambda}{n}\mathrm{pen}(\hat{f}_n) + \frac{\lambda}{n}\log(1/\delta),\\ &\leq \hat{r}_n(f',f) + \frac{\lambda}{n}\mathrm{pen}(f') + \frac{\lambda}{n}\log(1/\delta),\end{aligned} \tag{31}$$

with probability at least $1 - s\delta$, for all $f' \in \Gamma$. Applying Craig's Theorem once more, but this time to $\{-U_i\}_{i=1}^n$, using $\tau = \log(1/\delta)$, we get

$$\hat{r}_n(f',f) - r(f',f) < ar(f',f) + \frac{\lambda}{n}\log(1/\delta),$$

with probability at least $1-\delta$, therefore putting this together with (31) conclude that

$$(1-a)r(\hat{f}_n,f) < (1+a)r(f',f) + \frac{\lambda}{n}\mathrm{pen}(f') + \frac{2\lambda}{n}\log(1/\delta), \tag{32}$$

with probability at least $1-(s+1)\delta$. Or rearranging the various terms

$$r(\hat{f}_n,f) < \frac{1+a}{1-a}r(f',f) + \frac{\lambda}{n(1-a)}\mathrm{pen}(f') + \frac{2\lambda}{n(1-a)}\log(1/\delta), \tag{33}$$

with probability at least $1-(s+1)\delta$ for every $f' \in \Gamma$. Equation (33) is a *Probably Approximately Correct* (PAC) bound. It can easily be converted to an expected risk bound by a standard integration argument, using the fact that $\mathbb{E}[X] \leq \int_0^\infty \Pr(X \geq t)\mathrm{d}t$, for an arbitrary random variable X. To simplify the presentation let

$$\Upsilon(f',f) \equiv \frac{1+a}{1-a}r(f',f) + \frac{\lambda}{n(1-a)}\mathrm{pen}(f'),$$

and set $\delta = e^{-\frac{n(1-a)}{2\lambda}t}$. Then

$$\begin{aligned}\mathbb{E}\left[r(\hat{f}_n,f) - \Upsilon(f',f)\right] &\leq \int_0^\infty \Pr\left(r(\hat{f}_n,f) - \Upsilon(f',f) \geq t\right)\\ &\leq \int_0^\infty (s+1)e^{-\frac{n(1-a)}{2\lambda}t}\\ &= (s+1)\frac{2\lambda}{n(1-a)},\end{aligned}$$

for every $f' \in \Gamma$, yielding the final result. □

C Proof of Corollary 1

All there is to do is to check the moment condition for a Gaussian random variable. The moments of W_i are given by

$$\mathbb{E}[|W_i|^k] = \sigma^k \begin{cases} \prod_{i=1}^{k/2} (2i-1) & \text{, if } k \text{ is even,} \\ \sqrt{\frac{2}{\pi}} \prod_{i=1}^{(k-1)/2} (2i) & \text{, if } k \text{ is odd,} \end{cases}.$$

Using this fact one concludes that the moment condition is satisfied with $h = \frac{2}{3}\sqrt{\frac{2}{\pi}}\sigma$. Now by choosing the particular value of λ in the corollary statement we obtain the final result. □

D Proof of Theorem 8

Recall that a Hölder function with smoothness parameter α can be locally well approximated by a polynomial of degree $\lfloor \alpha \rfloor$, therefore it seems reasonable to consider piecewise polynomial estimators. Divide the domain $[0,1]^d$ in m^d disjoint hypercubes $\{U_{\boldsymbol{l}}\}_{\boldsymbol{l} \in \mathcal{L}}$, where

$$\mathcal{L} \equiv \{\boldsymbol{l} = (l_1, \ldots, l_d),\ l_i \in \{0, \ldots, m-1\}\} = \{0, \ldots, m-1\}^d,$$

and

$$U_{\boldsymbol{l}} = \left[\frac{l_1}{m}, \frac{l_1+1}{m}\right) \times \cdots \times \left[\frac{l_d}{m}, \frac{l_d+1}{m}\right) = U_{\boldsymbol{0}} + \frac{\boldsymbol{l}}{m},$$

with $\boldsymbol{l} = (l_1, \ldots, l_d) \in \mathcal{L}$. We will take m as an increasing function of n, so to be precise we should write m_n instead, but ease the notational burden we omit this dependence.

Let $T^{(i)}$, $i = 1, \ldots, D_\alpha$ be a basis for the space of polynomials of degree $\lfloor \alpha \rfloor$ over $\mathbb{R}^d$ (the dimension of this vector space is $D_\alpha \equiv \binom{\lfloor \alpha \rfloor + d}{\lfloor \alpha \rfloor}$). Furthermore assume that the restriction of $\{T^{(i)}\}$ to the set $U_{\boldsymbol{0}}$ is an orthogonal basis, that is, let $\{T^{(i)}\}$ satisfy

$$\int_{U_{\boldsymbol{0}}} T^{(i)}(\boldsymbol{x}) T^{(j)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0,$$

for $i \neq j$. Assume also that

$$\int_{U_{\boldsymbol{0}}} \left(T^{(i)}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} = M^2 \mathrm{vol}(U_{\boldsymbol{0}}) = \frac{M^2}{m^d},$$

where $\mathrm{vol}(\cdot)$ stands for the volume of a set.

In order to use Corollary 1 we need to construct a discrete class of possible estimators. To do so we consider a polynomial fit to each of the sets $U_{\boldsymbol{l}}$, $l \in \mathcal{L}$,

and quantize the coefficients of the polynomial representation with respect to a suitable basis. We consider estimates of the form

$$\gamma_{\boldsymbol{a}}(\boldsymbol{x}) \equiv \sum_{\boldsymbol{l}\in\mathcal{L}} \sum_{i=1}^{D_\alpha} a_{\boldsymbol{l}i}\ T^{(i)}\left(\boldsymbol{x} - \frac{\boldsymbol{l}}{m}\right) \mathbf{1}\{\boldsymbol{x} \in U_{\boldsymbol{l}}\},$$

where $\boldsymbol{a} \equiv \{a_{\boldsymbol{l}i}\}_{\boldsymbol{l}\in\mathcal{L}, i\in\{1,\ldots,D_\alpha\}}$. We need to consider a quantized version of the polynomial coefficients in order to have a discrete set of estimators. We will see later that it suffices to assume that

$$a_{\boldsymbol{l}j} \in \mathcal{Q}_n \equiv \{-1, \frac{-n+1}{n}, \ldots, \frac{n-1}{n}, 1\}.$$

Finally define

$$\Gamma = \left\{\gamma_{\boldsymbol{a}} : \boldsymbol{a} \in \mathcal{Q}_n^{\mathcal{L}\times\{1,\ldots,D_\alpha\}}\right\}.$$

Note that Γ is finite, and has $(2n+1)^{m^d D_\alpha}$ elements, therefore choosing $\mathrm{pen}(\gamma) = \log((2n+1)^{m^d D_\alpha})$ guarantees the Kraft inequality (14). Note also that, for $n \geq 2$, $\mathrm{pen}(\gamma) \leq 3D_\alpha m^d \log n$

We have now all the necessary ingredients to apply Corollary 1. Consider an arbitrary function $f \in \Sigma(L, \alpha, M)$. Notice that the penalty function is constant, therefore in the application of the oracle bound we just need to find an element of Γ that is sufficiently close to f, the function we want to estimate. Define $\bar{f} = \gamma_{\bar{\boldsymbol{a}}}$, where

$$\bar{\boldsymbol{a}} = \arg \min_{\boldsymbol{a}\in\mathbb{R}^{\mathcal{L}\times\{1,\ldots,D_\alpha\}}} \|f - \gamma_{\boldsymbol{a}}\|^2.$$

Since we chose an orthogonal basis, $\bar{\boldsymbol{a}} = \{\bar{a}_{\boldsymbol{l}i}\}_{\boldsymbol{l}\in\mathcal{L}, i\in\{1,\ldots,D_\alpha\}}$ is given by

$$\bar{a}_{\boldsymbol{l}i} = \frac{1}{M^2 \mathrm{vol}(U_{\boldsymbol{0}})} \int_{U_i} f(\boldsymbol{x}) T^{(i)}\left(\boldsymbol{x} - \frac{\boldsymbol{l}}{m}\right) \mathrm{d}\boldsymbol{x}.$$

Using the fact that $f \in \Sigma(L, \alpha)$ we have

$$\begin{aligned}
\|f - \bar{f}\|^2 &= \int_{[0,1]^d} |f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&= \sum_{\boldsymbol{l}\in\mathcal{L}} \int_{U_{\boldsymbol{l}}} |f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&\leq \sum_{\boldsymbol{l}\in\mathcal{L}} \int_{U_{\boldsymbol{l}}} |f(\boldsymbol{x}) - P_{\frac{\boldsymbol{l}}{m}}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \qquad (34) \\
&\leq \sum_{\boldsymbol{l}\in\mathcal{L}} \int_{U_{\boldsymbol{l}}} L^2 \left\|\boldsymbol{x} - \frac{\boldsymbol{l}}{m}\right\|^{2\alpha} \mathrm{d}\boldsymbol{x} \qquad (35) \\
&\leq \sum_{\boldsymbol{l}\in\mathcal{L}} \int_{U_{\boldsymbol{l}}} L^2 \left(\frac{d}{m^2}\right)^{\alpha} \mathrm{d}\boldsymbol{x} \\
&= \frac{L^2 d^\alpha}{m^{2\alpha}},
\end{aligned}$$

where (34) follows from the optimality of $\bar{f}$ and (35) follows from Definition 1 (recall that $P_{\frac{\boldsymbol{l}}{m}}(\cdot)$ is the Taylor polynomial of degree $\lfloor\alpha\rfloor$ around point $\boldsymbol{l}/m$).

Now observe that

$$\begin{aligned}
|\bar{a}_{\boldsymbol{l}i}| &= \left|\frac{1}{M^2\mathrm{vol}(U_{\mathbf{0}})}\int_{U_i} f(\boldsymbol{x})T^{(i)}\left(\boldsymbol{x}-\frac{\boldsymbol{l}}{m}\right)\mathrm{d}\boldsymbol{x}\right| \\
&\leq \frac{1}{M^2\mathrm{vol}(U_{\mathbf{0}})}\sqrt{\int_{U_i} f^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}\sqrt{\int_{U_i}\left(T^{(i)}\left(\boldsymbol{x}-\frac{\boldsymbol{l}}{m}\right)\right)^2\mathrm{d}\boldsymbol{x}} \\
&\leq \frac{1}{M^2\mathrm{vol}(U_{\mathbf{0}})}\sqrt{M^2\mathrm{vol}(U_{\mathbf{0}})}\sqrt{M^2\mathrm{vol}(U_{\mathbf{0}})} = 1. \qquad (36)
\end{aligned}$$

Define $f' = \gamma_{\boldsymbol{a}}$, where $a_{\boldsymbol{l}i} = \mathrm{round}(n\bar{a}_{\boldsymbol{l}i})/n$ (see footnote [10]). In light of the above we know that $a_{\boldsymbol{l}i} \in \mathcal{Q}_n$, therefore $f' \in \Gamma$. This is the estimate we are going to plug into the oracle bound (16). Begin by observing that

$$\begin{aligned}
\|\bar{f}-f'\|^2 &= \int_{[0,1]^d}|\bar{f}(\boldsymbol{x})-f'(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x} \\
&= \sum_{\boldsymbol{l}\in\mathcal{L}}\int_{U_{\boldsymbol{l}}}|\bar{f}(\boldsymbol{x})-f'(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x} \\
&= \sum_{\boldsymbol{l}\in\mathcal{L}}\int_{U_{\boldsymbol{l}}}\left|\sum_{i=1}^{D_\alpha}(\bar{a}_{\boldsymbol{l}i}-a_{\boldsymbol{l}i})T^{(i)}\left(\boldsymbol{x}-\frac{\boldsymbol{l}}{m}\right)\right|^2\mathrm{d}\boldsymbol{x} \\
&= \sum_{\boldsymbol{l}\in\mathcal{L}}\sum_{i=1}^{D_\alpha}(\bar{a}_{\boldsymbol{l}i}-a_{\boldsymbol{l}i})^2\int_{U_{\boldsymbol{l}}}T^{(i)}\left(\left(\boldsymbol{x}-\frac{\boldsymbol{l}}{m}\right)\right)^2\mathrm{d}\boldsymbol{x} \\
&= \sum_{\boldsymbol{l}\in\mathcal{L}}\sum_{i=1}^{D_\alpha}(\bar{a}_{\boldsymbol{l}i}-a_{\boldsymbol{l}i})^2M^2\mathrm{vol}(U_{\mathbf{0}}) \\
&\leq \sum_{\boldsymbol{l}\in\mathcal{L}}\sum_{i=1}^{D_\alpha}\frac{1}{4n^2}\frac{M^2}{m^d} = \frac{D_\alpha M^2}{4n^2}.
\end{aligned}$$

Putting these two bounds together we obtain

$$\begin{aligned}
\|f-f'\|^2 &\leq \|f-\bar{f}\|^2+\|\bar{f}-f'\|^2+2\|f-\bar{f}\|\|\bar{f}-f'\| \\
&\leq L^2d^\alpha\frac{1}{m^{2\alpha}}+\frac{D_\alpha M^2}{4}\frac{1}{n^2}+2\sqrt{\frac{L^2d^\alpha D_\alpha M^2}{4}}\frac{1}{nm^\alpha} \\
&= C\max\left\{\frac{1}{m^{2\alpha}},\frac{1}{n}\right\},
\end{aligned}$$

where $C>0$ is a constant.

Finally, using Corollary 1 we obtain

$$\mathbb{E}[\|f-\hat{f}_n\|] \leq C\max\left\{\frac{1}{m^{2\alpha}},\frac{1}{n},\frac{m^d\log n}{n}\right\},$$

[10] Let $x\in\mathbb{R}$, define $\mathrm{round}(x)\equiv\arg\min_{z\in\mathbb{Z}}|x-z|$. Clearly $|\mathrm{round}(x)-x|\leq 1/2$.

for a suitable $C > 0$, and any $n \geq 2$, therefore choosing $m = \left\lceil (n/\log n)^{\frac{1}{2\alpha+d}} \right\rceil$ yields

$$\mathbb{E}[\|f - \hat{f}_n\|] \leq C(n/\log n)^{-\frac{2\alpha}{2\alpha+d}},$$

for some $C \equiv C(L, \alpha, M, \sigma^2)$ and any $n \geq 2$. □

E Proof of Theorem 9

As mentioned before, we are going to analyze a modification of the estimator described in (21). This proof uses many of the ideas introduced in the proof of Theorem 8, therefore it is recommended that the reader glances at that proof before proceeding this section.

Let $\pi = \{A_1, \ldots, A_k\} \in \Pi$ be a RDP and let $\{T_{A_l}^{(i)}\}_{i=1}^{D_\alpha}$, $l = 1, \ldots, k$, be an orthogonal basis of the space of polynomials of degree $\lfloor \alpha \rfloor$ over A_k. In particular, for any $l \in \{1, \ldots, k\}$ we require the basis to have the following properties

i) For all $i, j \in \{1, \ldots, D_\alpha\}$, $i \neq j$

$$\int_{A_l} T_A^{(i)}(\boldsymbol{x}) T_A^{(j)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0, \ \forall A \in \pi, \text{ and}$$

ii)

$$\int_A \left(T_A^{(i)}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} = M^2 \mathrm{vol}(A), \ \forall A \in \pi.$$

We now define the discrete analogue of $\Xi(\pi)$ (see equation (19)). We do this by describing the polynomials decoration each set of the partition using the previously constructed basis, and restricting the coefficients of those representations to lie on the discrete set

$$\mathcal{Q}_n \equiv \{-1, \frac{-n+1}{n}, \ldots, \frac{n-1}{n}, 1\}.$$

Therefore we define

$$\Xi_{\mathcal{Q}_n}(\pi) = \left\{ \sum_{A \in \pi} \sum_{i=1}^{D_\alpha} a_{A,i} \, T_A^{(i)}(\boldsymbol{x}) \mathbf{1}\{\boldsymbol{x} \in A\} : \ a_{A,i} \in \mathcal{Q}_n \ \forall A, i \right\},$$

and consequently the class of possible estimators we consider is

$$\Gamma = \bigcup_{\pi \in \Pi} \Xi_{\mathcal{Q}_n}(\pi). \tag{37}$$

This is clearly a countable set (although not finite), and is the set of estimates we will use to apply Corollary 1.

To construct a penalty function that satisfies the Kraft inequality (14) we use an explicit description of a prefix encoding of the elements of Γ, therefore automatically satisfying (14). Let $\gamma_\pi \in \Xi_{\mathcal{Q}_n}(\pi) \subseteq \Gamma$. The encoding of an element

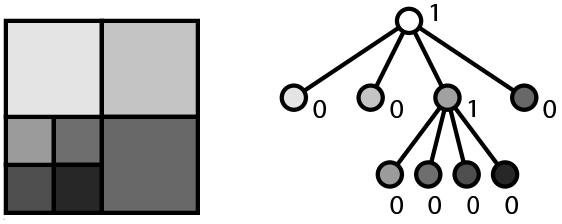


Figure 5: Prefix encoding of a Recursive Dyadic Partition. The depicted partition encodes as 100100000 in binary.

of γ_π is done in two steps: (i) encoding the underlying RDP π, (ii) encoding the coefficients of the decorating polynomials $\{a_{Ai}\}_{A\in\pi, i\in\{1,\dots,D_\alpha\}}$. To encode the underlying RDP we resort to its tree representation (refer to Figure 5), and assign a zero or one value to each node of the tree: zero if that node is a leaf node, and one otherwise. Now collect all those values in a lexicographical order,*i.e.*left-to-right breadth-first order (see example in Figure 5). This forms a binary prefix code for that space of RDP trees.

Note that each node in a RDP tree has either zero or 2^d descendants, therefore the tree has $1 + 2^d k$ nodes, for some $k \in \mathbb{N}_0$, and it has $1 + (2^d - 1)k$ leaf nodes. The number of leaf nodes is the size of the RDP and so, we can describe a RDP π using

$$\lambda_1(\pi) \equiv 1 + \frac{2^d}{2^d - 1}(|\pi| - 1)$$

bits. Notice that since this is a binary prefix code it satisfies the Kraft inequality $\sum_{\pi\in\Pi} 2^{-\lambda_1(\pi)} \leq 1$.

For each element of the RDP we have a polynomial with D_α coefficients taking values over $\mathcal{Q}_n$, therefore $\Xi_{\mathcal{Q}_n}$ has $(2n+1)^{|\pi|D_\alpha}$ elements. With this at hand define

$$\mathrm{pen}(\gamma_\pi) = \left(\frac{2^d \log 2}{2^d - 1} + D_\alpha \log(2n+1)\right)|\pi|. \tag{38}$$

We have the following result

Lemma 2. *The penalty defined in* (38) *satisfies* (14)*, that is*

$$\sum_{\gamma\in\Gamma} \exp(-\mathrm{pen}(\gamma)) \leq 1$$

for Γ *defined in* (37)*.*

Proof.

$$\begin{aligned}
\sum_{\gamma\in\Gamma}\exp(-\text{pen}(\gamma)) &= \sum_{\bigcup_{\pi\in\Pi}\Xi_{\mathcal{Q}_n}(\pi)}\exp(-\text{pen}(\gamma)) \\
&\leq \sum_{\pi\in\Pi}\sum_{\gamma\in\Xi_{\mathcal{Q}_n}(\pi)}\exp\left(-\frac{2^d\log 2}{2^d-1}|\pi| - D_\alpha\log(2n+1)|\pi|\right) \\
&= \sum_{\pi\in\Pi}\exp\left(-\frac{2^d\log 2}{2^d-1}|\pi|\right)\sum_{\gamma\in\Xi_{\mathcal{Q}_n}(\pi)}\frac{1}{(2n+1)^{|\pi|D_\alpha}} \\
&= \sum_{\pi\in\Pi}2^{-\frac{2^d}{2^d-1}|\pi|} \\
&\leq \sum_{\pi\in\Pi}2^{-\left(1+\frac{2^d}{2^d-1}(|\pi|-1)\right)} = \sum_{\pi\in\Pi}2^{-\lambda_1(\pi)} \leq 1.
\end{aligned}$$

□

It's important to note that $\text{pen}(\gamma_\pi) \leq C\log n|\pi|$ for any $n \geq 2$ and a suitable constant $C > 0$. This is going to be used later to avoid dealing with cumbersome constants.

We are now ready to apply Corollary 1. We consider first the case $d > 1$. Fix an arbitrary function $f \in \mathcal{F}(L,\alpha,\beta,M)$. Our strategy is to construct a partition that is well adapted to the boundary set $B(f)$, in the sense that the partition sets that intersect $B(f)$ are small. This is desirable because the discontinuity of $B(f)$ cannot be well approximated with polynomials. Away from the boundary we can use larger partition sets, although not too large, so that polynomials approximate well the function f over those sets.

Let $J, J' \in \mathbb{N}_0$ and $J > J'$. Consider the RDP tree with all the leafs at depth J. The corresponding RDP has 2^{dJ} elements. Now prune this tree so that leafs intersecting $B(f)$ are at depth J and all the other leafs are between depths J' and J. This process is illustrated in Figure 6. We have the following result:

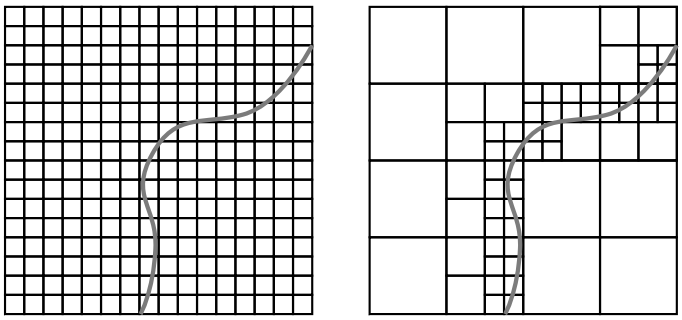
Lemma 3. *There is a RDP such that leafs intersecting $B(f)$ are at depth J and all the other leafs are between depths J' and J ($J' < J$). Denote the smallest such RDP by $\pi_{J',J}$. This RDP has at most $2^{2d}B2^{(d-1)J}$ leafs intersecting $B(f)$ and*

$$|\pi_{J',J}| \leq 2^{d(J'+1)} + \begin{cases} B'J & \text{, if } d = 1 \\ B'2^{(d-1)J} & \text{, if } d > 1 \end{cases},$$

where

$$B' = \begin{cases} 2^{2d}B & \text{, if } d = 1 \\ \frac{2^{3d-1}}{2^{d-1}-1}B & \text{, if } d > 1 \end{cases}.$$

Proof. Denote the smallest RDP satisfying the conditions of the lemma by $\pi_{J,J'}$. The number of leafs is trivially bounded by the number of nodes in the tree,



(a) (b)

Figure 6: Example of RDP tree pruning, for $d = 2$, $J = 4$, and $J' = 2$. The depicted curve is $B(f)$: (a) partition with all leafs at depth J; (b) pruned partition adapted to $B(f)$.

and so the proof strategy entails by bounding from above the number of nodes $\pi_{J,J'}$ might have. First, since the all the leafs are at depth greater of equal than J', $\pi_{J,J'}$ has no more than $2^{d(J'+1)}$ nodes at depth J'. Now let $j \in \mathbb{N}$. Begin by noticing that any closed ball of diameter 2^{-j} is contained in at most 2^d nodes at depth j, thus at depth j there are at most $2^d B 2^{(d-1)j}$ nodes of $\pi_{J,J'}$ that intersect $B(f)$ (recall Definition 3). Due to the diadic structure every leaf has $2^d - 1$ siblings, thus $\pi_{J,J'}$ has less than $2^{2d} B 2^{(d-1)j}$ nodes at depth j (therefore $\pi_{J',J}$ has at most $2^{2d} B 2^{(d-1)J}$ leafs). Finally, the total number of nodes of $\pi_{J,J'}$ is bounded from above by

$$2^{d(J'+1)} + 2^{2d} B \sum_{j=J'+1}^{J} 2^{(d-1)j} \tag{39}$$

$$\leq \quad 2^{d(J'+1)} + 2^{2d} B \sum_{j=0}^{J} 2^{(d-1)j} \tag{40}$$

$$\leq \quad 2^{d(J'+1)} + \begin{cases} 2^{2d} BJ & \text{, if } d = 1 \\ \frac{2^{3d-1}}{2^{d-1}-1} B 2^{(d-1)J} & \text{, if } d > 1 \end{cases} . \tag{41}$$

□

The key point of Lemma 3 is that the total number of leafs in the described tree is of the same order of magnitude has the number of leafs intersecting the boundary set (for $d > 1$).

Let $\pi_{J',J}$ be the partition of Lemma 3 and define

$$\bar{f} = \sum_{A \in \pi} \sum_{i=1}^{D_\alpha} \bar{a}_{A,i} \, T_A^{(i)}(\boldsymbol{x}) \mathbf{1}\{\boldsymbol{x} \in A\},$$

where

$$\bar{a}_{Ai} = \frac{1}{M^2 \mathrm{vol}(A))} \int_A f(\boldsymbol{x}) T^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

It is easy to verify that $\bar{f}$ minimizes $\|f - \bar{f}\|^2$ over $\Xi(\pi_{J',J})$. Also, by the same reasoning used in (36)

$$|\bar{a}_{Ai}| \leq 1, \ \forall A \in \pi, i \in \{1, \ldots, D_\alpha\}.$$

Finally define

$$f' = \sum_{A \in \pi} \sum_{i=1}^{D_\alpha} a_{A,i} \ T_A^{(i)}(\boldsymbol{x}) \mathbf{1}\{\boldsymbol{x} \in A\},$$

where $a_{A,i} = \mathrm{round}(n\bar{a}_{A,i})/n$. Clearly $a_{A,i} \in \mathcal{Q}_n$, and so $f' \in \Gamma$. This is the estimate we are going to plug-in the oracle bound (16).

As in the proof of Theorem 8 we need to bound $\|f - f'\|$. We consider first the case $d > 1$. For the ease on notation define two subsets of π: $\pi_1 \equiv \{A \in \pi : A \bigcap B(f) \neq \emptyset\}$, the set of elements of π that intersect $B(f)$, and $\pi_2 \equiv \pi \setminus \pi_1$, the set of elements of π that do not intersect $B(f)$.

$$\begin{aligned}
\|f - \bar{f}\|^2 &= \int_{[0,1]^d} |f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&= \sum_{A \in \pi} \int_A |f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&= \sum_{A \in \pi_1} \int_A |f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} + \sum_{A \in \pi_2} \int_A |f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&\leq \sum_{A \in \pi_1} M^2 \mathrm{vol}(A) + \sum_{A \in \pi_2} \int_A |f(\boldsymbol{x}) - P_{\boldsymbol{z}_A}|^2 \mathrm{d}\boldsymbol{x} \qquad (42) \\
&\leq B' 2^{(d-1)J} M^2 2^{-dJ} + \sum_{A \in \pi_2} \int_A L^2 |\boldsymbol{x} - \boldsymbol{z}_A|^{2\alpha} \mathrm{d}\boldsymbol{x} \\
&\leq B' M^2 2^{-J} + \sum_{A \in \pi_2} \int_A L^2 \mathrm{diam}(A)^{2\alpha} \mathrm{d}\boldsymbol{x} \\
&= B' M^2 2^{-J} + \sum_{A \in \pi_2} L^2 (d 2^{-2J'})^\alpha \mathrm{vol}(A) \\
&\leq B' M^2 2^{-J} + L^2 d^\alpha 2^{-2\alpha J'}, \qquad (43)
\end{aligned}$$

where $\mathrm{diam}(\cdot)$ stands for diameter of a set, and step (42) is due to the fact that $\bar{f}$ minimizes $\|f - \bar{f}\|^2$ over $\Xi(\pi_{J',J})$. Also

$$\begin{aligned}
\|\bar{f} - f'\|^2 &= \int_{[0,1]^d} |\bar{f}(\boldsymbol{x}) - f'(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&= \sum_{A \in \pi} \int_A |\bar{f}(\boldsymbol{x}) - f'(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{A\in\pi}\int_A |\sum_{i=1}^{D_\alpha}(\bar{a}_{A,i}-a_{A,i})T_A^{(i)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&= \sum_{A\in\pi}\sum_{i=1}^{D_\alpha}(\bar{a}_{A,i}-a_{A,i})^2\int_A \left(T_A^{(i)}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} \\
&= \sum_{A\in\pi}\sum_{i=1}^{D_\alpha}(\bar{a}_{A,i}-a_{A,i})^2 M^2 \mathrm{vol}(A) \\
&\leq \sum_{A\in\pi}\sum_{i=1}^{D_\alpha}\frac{1}{4n^2}M^2\mathrm{vol}(A) = \frac{D_\alpha M^2}{4n^2}.
\end{aligned}
$$

Finally

$$
\begin{aligned}
\|f-f'\|^2 &\leq \|f-\bar{f}\|^2 + \|\bar{f}-f'\|^2 + 2\|f-\bar{f}\|\|\bar{f}-f'\| \\
&\leq C\max\left\{2^{-J}, 2^{-2\alpha J'}, \frac{1}{n}\right\},
\end{aligned}
$$

where $C>0$ is a suitable constant. Then Corollary 1 yields

$$
\mathbb{E}[\|f-\hat{f}_n\|^2] \leq C\max\left\{2^{-J}, 2^{-2\alpha J'}, \frac{1}{n}, \frac{2^{dJ'}\log n}{n}, \frac{2^{(d-1)J}\log n}{n}\right\},
$$

for a suitable $C>0$ and any $n\geq 2$, where the two last arguments in the maximum function follow from Lemma 3. Choosing $J=\lceil \frac{1}{d}\log(n/\log(n))\rceil$ and $J'=\lceil \frac{1}{2\alpha+d}\log(n/\log(n))\rceil$ yields the desired result

$$
\mathbb{E}[\|f-\hat{f}_n\|^2] \leq C\max\left\{\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}}, \left(\frac{n}{\log n}\right)^{-\frac{1}{d}}\right\},
$$

for some $C\equiv C(L,\alpha,\beta,M,\sigma^2)$ and all $n\geq 2$.

When $d=1$ the argument suffers a slight modification. Instead of (43) we have

$$
\|f-\bar{f}\|^2 \leq B'M^2J2^{-J} + L^2 2^{-2\alpha J'},
$$

which follows by the same reasoning as before but noting that Lemma 3 gives a different expression for $|\pi_1|$. From Corollary 1

$$
\mathbb{E}[\|f-\hat{f}_n\|^2] \leq C\max\left\{J2^{-J}, 2^{-2\alpha J'}, \frac{1}{n}, \frac{2^{J'}\log n}{n}, \frac{J\log n}{n}\right\}.
$$

With $J=\lceil \log n\rceil$ and J' as before we obtain

$$
\begin{aligned}
\mathbb{E}[\|f-\hat{f}_n\|^2] &\leq C'\max\left\{\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}}, \frac{\log^2 n}{n}\right\} \\
&\leq C\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}},
\end{aligned}
$$

for some $C'>0$, $C\equiv C(L,\alpha,\beta,M,\sigma^2)$ and all $n\geq 2$, concluding the proof. □

F Sketch of the proof of Theorem 10

This proof is a simplified version of the proof of Theorem 9. Using the notation introduced in that proof, the modifications are as follows. Consider first the case $d > 1$. The estimators we consider are obtained decorating each leaf on an RDP with a constant (instead of a polynomial of degree $\lfloor \alpha \rfloor$). Since the functions in the class under consideration are piecewise constant, this model is exact away from the boundary and so we get

$$\|f - \bar{f}\|^2 \leq B' M^2 2^{-J},$$

and

$$\|\bar{f} - f'\|^2 \leq \frac{M^2}{4n^2}.$$

Putting these results together and using Corollary 1 in the same fashion as before we get

$$\mathbb{E}[\|f - \hat{f}_n\|^2] \leq C \max\left\{2^{-J}, \frac{1}{n}, \frac{2^{(d-1)J}\log n}{n}\right\},$$

for a suitable $C > 0$ and any $n \geq 2$. Choosing $J = \lceil \frac{1}{d}\log(n/\log(n)) \rceil$ yields the desired result

$$\mathbb{E}[\|f - \hat{f}_n\|^2] \leq C\left(\frac{n}{\log n}\right)^{-\frac{1}{d}},$$

for some $C \equiv C(\beta, M, \sigma^2)$ and all $n \geq 2$.

When $d = 1$ we have

$$\|f - \bar{f}\|^2 \leq B' M^2 J 2^{-J},$$

which follows by the same reasoning as before but noting that Lemma 3 gives a different expression for $|\pi_1|$. Corollary 1 yields

$$\mathbb{E}[\|f - \hat{f}_n\|^2] \leq C \max\left\{J2^{-J}, \frac{1}{n}, \frac{J\log n}{n}\right\}.$$

Choosing $J = \lceil \log n \rceil$ we obtain

$$\mathbb{E}[\|f - \hat{f}_n\|^2] \quad \leq \quad C\frac{\log^2 n}{n},$$

for some $C \equiv C(\beta, M, \sigma^2)$ and all $n \geq 2$, concluding the proof. □

G Proof of Theorem 11

Consider the partition of $[0,1]^d$ into m^d identical hypercubes, just as in statement of (A4). Let $m = 2^J$ and $J \in \mathbb{N}$. This partition, that we denote by π_J, can also be constructed with a RDP where all the leaves are at depth J. Using

this partition we define $f_J : [0,1]^d \to \mathbb{R}$, a coarse approximation of f up to resolution J. Formally we have

$$f_J(\boldsymbol{x}) = \frac{1}{m^d} \sum_{A \in \pi_J} \left(\int_A f(\boldsymbol{t}) \mathrm{d}\boldsymbol{t} \right) \mathbf{1}_A(\boldsymbol{x}).$$

Note that f_J is identical to f "away" from the boundary (since f is piecewise constant), but in the vicinity of the boundary there is some averaging. We have the following important result.

Lemma 4. *Let $\hat{f}_l^p$ be the complexity regularized estimators introduced above (with $l \in \{0, \ldots, d\}$). Let $f \in BF(M)$. Then*

$$\mathbb{E}\left[\|\hat{f}_l^p - f_J\|^2 \right] \leq C \frac{2^{(d-1)J} \log n'}{n'},$$

for a suitable $C > 0$, and all $n' \geq 2$.

Proof. The key idea used in the proof is the constructing of a modified observation setup, that is, instead of using $\{\boldsymbol{X}_i, Y_i\}_{i=1}^{n'}$ to determine the estimator $\hat{f}_l^P$, we use a different observation model, yielding observations $\{\boldsymbol{X'}_i, Y_i'\}_{i=1}^{n'}$. This new observation model is carefully chosen so that the output to the estimator using either $\{\boldsymbol{X}_i, Y_i\}_{i=1}^{n'}$ or $\{\boldsymbol{X'}_i, Y_i'\}_{i=1}^{n'}$ is statistically indistinguishable.

The new observation model is of the form

$$Y_i = f_J(\boldsymbol{X}_i) + W_i',$$

where $\{W_i\}$ are all independent but not identically distributed. Namely let $A_{\boldsymbol{X}_i}$ denote the partition set where $\boldsymbol{X}_i$ is contained, that is,

$$A_{\boldsymbol{X}_i} \equiv A : A \in \pi_J \text{ and } \boldsymbol{X}_i \in A.$$

We define

$$W_i' \equiv f(\boldsymbol{U}_i) - f_J(\boldsymbol{U}_i) + W_i,$$

where $\{\boldsymbol{U}_i\}_{i=1}^{n'}$ are all independent of $\{W_i\}$ and $\boldsymbol{U}_i | \boldsymbol{X}_i \sim \mathrm{Unif}(A_{\boldsymbol{X}_i})$, and $\{W_i\}_{i=1}^{n'}$ are (i.i.d.) Gaussian with zero mean and variance σ^2.

Notice that the estimators $\hat{f}_l^P$ average the data within each partition cell $A \in \pi_J$, completely ignoring the sample location $\boldsymbol{X}_i$ within the cell. This ensures that the above observation model is statistically indistinguishable from the original observation model, when used by the estimation procedure. Note that under the new observation model the regression function is $\mathbb{E}[Y_i | \boldsymbol{X}_i = \boldsymbol{x}]$ is $f_J(\boldsymbol{x})$, instead of $f(\boldsymbol{x})$ for the original observation model. This is the key to obtain the desired result, following from the application of Theorem 7, since now we can evaluate the error performance with respect to f_J. We just need to check that W_i' satisfies the moment condition. Since W_i' is the sum of two independent random variables, namely W_i and $f(\boldsymbol{U}_i) - f_J(\boldsymbol{U}_i)$ we can again use the result in [11] for the sum of random variables satisfying the moment

condition (as in the proof of Theorem 7). Therefore W_i' satisfies the moment condition (13) with constant $h = 8(\frac{2}{3}\sqrt{\frac{2}{\pi}} + \frac{4M^2}{3})$. From here we proceed as in the proof of Theorem 8, by noting that there is a model in Γ, built over a partition with $2^{(d-1)J}$ elements, that approximates f_J extremely well. Therefore we conclude that

$$\mathbb{E}\left[\|\hat{f}_l^p - f_J\|^2\right] \leq C \max\left\{\frac{2^{(d-1)J}\log n'}{n'}, \frac{1}{n'}\right\} = C\frac{2^{(d-1)J}\log n'}{n'},$$

for a suitable $C > 0$, and all $n' \geq 2$. □

To bound the risk of the active learning procedure we are going to consider the error incurred in three different situations: (i) the error incurred during the preview stage in regions away from the boundary; (ii) the error incurred by not detecting a piece of the boundary (and therefore not performing the refinement stage on that area); (iii) the error incurred during the refinement stage.

(i) - Recall that f_J is identical to f "away" from the boundary set $B(f)$. That is, for a fixed set $A \in \pi_J$ that does not intersect the boundary we have $f(\boldsymbol{x}) = f_J(\boldsymbol{x})$ for all $\boldsymbol{x} \in A$. Therefore Lemma 4 characterizes the behavior of the preview estimator "away" from the boundary. Let $\mathcal{I}$ denote the union of all partition sets of π_J not intersecting the boundary. Then

$$\begin{aligned}\mathbb{E}\left[\int_{\mathcal{I}} |\hat{f}_{\text{active}}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}\right] &= \mathbb{E}\left[\int_{\mathcal{I}} |\hat{f}_0^p(\boldsymbol{x}) - f_J(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}\right] \\ &\leq \mathbb{E}\left[\|f_0^p - f_J\|^2\right] \\ &\leq C\frac{2^{(d-1)J}\log n'}{n'},\end{aligned}$$

for a suitable $C > 0$ and all $n' \geq 2$.

(ii) - Let $\hat{\mathcal{I}}$ be the set of elements of π_J intersecting the boundary that are not going to be re-sampled in the refinement step. That is

$$\hat{\mathcal{I}} = \left\{A \in \pi_J, A \cap B(f) \neq \emptyset : \forall l \in \{0, \ldots, d\}\ A \notin \hat{\pi}_l^p\right\}.$$

Under (A4) we know that for each element of $\hat{\mathcal{I}}$ (24) holds for at least one of the shifted RDPs. Therefore we can construct a decomposition

$$\hat{\mathcal{I}} = \hat{\mathcal{J}}_0 \cup \hat{\mathcal{J}}_1 \cup \cdots \cup \hat{\mathcal{J}}_d,$$

where $\hat{\mathcal{J}}_l$ are disjoint and (24) holds for all the elements of $\hat{\mathcal{J}}_l$, with respect to shifted RDPs in the l^{th} coordinate (Notice that many such decompositions might exist, but for our purposes we just need to consider one of

them). Now

$$\begin{aligned}
\mathbb{E}[\|\hat{f}_l^p - f_J\|^2] &\geq \mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{J}}_l}\mathcal{A}(A,l)} \left(\hat{f}_l^p(\boldsymbol{x}) - f_J(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}\right] \\
&= \mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{J}}_l}\mathcal{A}(A,l)} \left(f_J(\boldsymbol{x}) - \mathbb{E}[\hat{f}_l^p(\boldsymbol{x})|\hat{\pi}_l^p]\right)^2 \mathrm{d}\boldsymbol{x}\right] + \\
&\quad \mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{J}}_l}\mathcal{A}(A,l)} \left(\hat{f}_l^p(\boldsymbol{x}) - \mathbb{E}[\hat{f}_l^p(\boldsymbol{x})|\hat{\pi}_l^p]\right)^2 \mathrm{d}\boldsymbol{x}\right] \\
&\geq \mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{J}}_l}\mathcal{A}(A,l)} \left(f_J(\boldsymbol{x}) - \mathbb{E}[\hat{f}_l^p(\boldsymbol{x})|\hat{\pi}_l^p]\right)^2 \mathrm{d}\boldsymbol{x}\right] \\
&\geq \mathbb{E}[|\hat{\mathcal{J}}_l|]C_b(f)2^{-dJ}.
\end{aligned}$$

Using Lemma 4 we conclude that $\mathbb{E}[|\hat{\mathcal{J}}_l|] \leq \frac{2^{(2d-1)J}\log n'}{n'}\frac{1}{C_b(f)}$, and therefore $\mathbb{E}[|\hat{\mathcal{I}}|] \leq (d+1)\frac{2^{(2d-1)J}\log n'}{n'}\frac{1}{C_b(f)}$. The maximum error we incur in our final estimate by erroneously not detecting certain pieces of the boundary is bounded above by

$$\begin{aligned}
&\mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{I}}} A} \left(\hat{f}_{\text{active}}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}\right] \\
&= \mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{I}}} A} \left(\hat{f}_0^p(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}\right] \\
&= \mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{I}}} A} \left(f(\boldsymbol{x}) - \mathbb{E}[\hat{f}_0^p(\boldsymbol{x})]\right)^2 \mathrm{d}\boldsymbol{x}\right] + \\
&\quad \mathbb{E}\left[\int_{\bigcup_{A\in\hat{\mathcal{I}}} A} \left(\hat{f}_0^p(\boldsymbol{x}) - \mathbb{E}[\hat{f}_0^p(\boldsymbol{x})]\right)^2 \mathrm{d}\boldsymbol{x}\right] \\
&\leq \mathbb{E}[|\hat{\mathcal{I}}|]M^2 2^{-dJ} + C\frac{2^{(d-1)J}\log n'}{n'} \leq C'\frac{2^{(d-1)J}\log n'}{n'},
\end{aligned}$$

Where $C' > 0$, and $C > 0$ comes from Lemma 4. We conclude that the error incurred by failing to detect the boundary has the same contribution for the total error of the estimator as the error away from the boundary, analyzed in (i).

(iii) - In the regions that are going to be refined, that is, the regions in $\mathcal{R}$, we are going to collect further samples and apply the estimator described in Section 4.1. Assume for now that we have $O(2^{(d-1)J})$ elements in $\mathcal{R}$. This is going to be proved later on, in Lemma 5. We collect a total of $L \equiv n'/|\mathcal{R}|$ samples in each element of $\mathcal{R}$. The error incurred by $\hat{f}_r$, the refinement estimator, over each one of the elements of $\mathcal{R}$ is upper-bounded by

$$C\left(\frac{\log L}{L}\right)^{1/d} 2^{-dJ},$$

where $C > 0$ comes from Theorem 10. Therefore the error of the estimator over $\cup_{A\in\mathcal{R}} A$ is upper-bounded by

$$C\left(\frac{\log L}{L}\right)^{1/d} 2^{-dJ}|\mathcal{R}|.$$

To compute the total error incurred by $\hat{f}_{\text{active}}$ we just have to sum the contributions of (i), (ii) and (iii), and therefore we get

$$\mathbb{E}\left[\|\hat{f}_{\text{active}} - f\|^2\right] \leq C\left(\frac{\log L}{L}\right)^{1/d} 2^{-dJ}|\mathcal{R}| + C'\frac{2^{(d-1)J}\log n'}{n'},$$

with $C, C' > 0$. Assuming now that $|\mathcal{R}| = O(2^{(d-1)J})$ we can balance the two terms in the above expression by choosing

$$J = \left\lceil \frac{d-1}{(d-1)^2 + d} \log(n'/\log(n')) \right\rceil,$$

yielding the desired result.

As mentioned before, we need to show that the number of partition sets where we are going to distribute samples in the refinement step is not very large, namely, with high probability $\#\mathcal{R} = O(2^{(d-1)J})$. This ensures that there are enough samples in each one of these partition sets to properly perform the refinement step. Without loss of generality, it suffices to analyze the number of elements in $\hat{\pi}_0^p$. We will denote this estimator by $\hat{\pi}$ to ease the notation.

Lemma 5. *Let $\hat{\pi}$ be the partition estimated according to* (21)*. Let $\pi_{J',J}$ be the partition adapted to the boundary, according to Lemma 3 (recall that this cannot computed from the data). Then with high probability the number of elements of $\hat{\pi}$ is comparable with the number of elements of $\pi_{J',J}$, namely*

$$\Pr(|\hat{\pi}| > 2|\pi_{J',J}|) \leq 1/n,$$

for n sufficiently large.

From this lemma we conclude that, with high probability, the number of cells to be refined is actually $O(2^{(d-1)J})$ and so all the analysis done before holds, with probability $1 - (1/n)$, concluding the proof.

Proof of Lemma 5: First note that the number of elements in $\hat{\pi}_0^p$ is equal to $(2^d-1)k+1$ for $k \in \mathbb{N}_0$, due to the dyadic structure of the partitions. Our goal is to bound

$$\begin{aligned}\Pr\left(|\hat{\pi}_0^p| = (2^d-1)k+1\right) &= \Pr\left(\bigcup_{\pi:|\pi|=(2^d-1)k+1} \{\hat{\pi}=\pi\}\right) \\ &\leq \sum_{\pi:|\pi|=(2^d-1)k+1} \Pr\left(\hat{\pi}=\pi\right) \\ &\leq \#_k \max_{\pi:|\pi|=(2^d-1)k+1} \left\{\Pr\left(\hat{\pi}=\pi\right)\right\}, \qquad (44)\end{aligned}$$

where $\#_k$ is the number of partitions with $(2^d-1)k+1$ elements. A very crude upper-bound on $\#_k$ is $\binom{2^{dJ}}{k}$. This is obtained noticing that an RDP with $(2^d-1)k+1$ elements is constructed by doing k splits of the trivial RDP (as in the formal rules for the construction of RDPs).

To bound $\Pr(\hat{\pi}=\pi)$ recall Lemma 3. Let π be an arbitrary RDP. There is another partition π' that can be constructed from π by aggregation, adapted to the boundary and such that

$$|\pi'| \leq \min\left(|\pi|, (2^d-1)C2^{(d-1)J}+1\right),$$

where $C>0$ comes from Lemma 3. If $k \leq C2^{(d-1)J}$ we upper bound $\Pr(\hat{\pi}=\pi)$ by trivially by one. If $k > C2^{(d-1)J}$ notice that π and π' are nested and $\pi \preceq \pi'$. To bound $\Pr(\hat{\pi}=\pi)$ we will bound the probability that the estimation strategy chooses π against π'. For a fixed partition the model fit corresponds simply to a projection onto a linear space. Recall (21). We choose π against π' if the difference between the squared errors of the model fits for π and π' is greater then the difference of the respective penalty terms. Noting again that $\pi \preceq \pi'$ (that is π is nested inside π') the difference between the squared errors is a χ^2 random variable, with $|\pi|-|\pi'|$ degrees of freedom, and so

$$\Pr(\hat{\pi}=\pi) \leq \Pr\left(U_{(2^d-1)(k-C2^{(d-1)J})} > 2p(n)(2^d-1)(k-C2^{(d-1)J})\right),$$

where $U_{(2^d-1)(k-C2^{(d-1)J})}$ is a χ^2 random variable with $(2^d-1)(k-C2^{(d-1)J})$ degrees of freedom, and $p(n)=c_0\log n$. In [15] Laurent and Massart state the following lemma (Lemma 1): If U_q is χ^2 distributed with q degrees of freedom then, for $s>0$

$$\Pr(U_q \geq q + s\sqrt{2q} + s^2) \leq e^{-s^2/2}.$$

Now take $q=(2^d-1)(k-C2^{(d-1)J})$ and $q+s\sqrt{2q}+s^2 = 2p(n)q$. After some manipulation we conclude that

$$\begin{aligned}\Pr(\hat{\pi}=\pi) &\leq \exp\left(-q\left(p(n)-\sqrt{p(n)-1/4}\right)\right) \\ &= \exp\left(-(2^d-1)(k-C2^{(d-1)J})\left(p(n)-\sqrt{p(n)-1/4}\right)\right).\end{aligned}$$

We can now ask for a bound on the probability that the number of elements of $\hat{\pi}$ exceeds some value. In particular we are going to bound the probability that the chosen partition has approximately twice more leafs than the optimal partition, adapted clairvoyantly to the boundary set. Concretely, we are going to bound

$$\zeta \equiv \Pr(|\hat{\pi}| \geq (2^d - 1)2C2^{(d-1)J} + 1).$$

Using (44) we have

$$\zeta \leq \sum_{k=2C2^{(d-1)J}}^{\infty} \left\{ \binom{2^{dJ}}{k} \exp\left(-(2^d-1)(k - C2^{(d-1)J})\left(c_0 \log n - \sqrt{c_0 \log n - 1/4}\right)\right)\right\}.$$

Let $M = 2^{dJ}$ and note that $p(n) = c_1 \log M$. Then

$$\zeta \leq \sum_{k=2CM^{\frac{d-1}{d}}}^{\infty} \left\{ \binom{M}{k} \exp\left(-(2^d-1)(k - CM^{\frac{d-1}{d}})\left(c_1 \log M - \sqrt{c_1 \log M - 1/4}\right)\right)\right\}.$$

For M large the $\log M$ term dominates the $\sqrt{\log M}$ term, and so, for $\epsilon > 0$ and M sufficiently large we have

$$\begin{aligned}
\zeta &\leq \sum_{k=2CM^{\frac{d-1}{d}}}^{\infty} \binom{M}{k} \exp\left(-(2^d-1)(k - CM^{\frac{d-1}{d}})c_1 \log M(1-\epsilon)\right) \\
&\leq \sum_{k=2CM^{\frac{d-1}{d}}}^{\infty} \frac{M^k}{k!} M^{-(2^d-1)(k-CM^{\frac{d-1}{d}})c_1(1-\epsilon)}.
\end{aligned}$$

Now we use the fact that $k!$ grows much faster than an exponential, namely, for M sufficiently large we have $k! > M^{\alpha k}$ for some $\alpha > 0$. Take α such that $1 - \alpha - (2^d - 1)c_1(1-\epsilon) < 0$. Then, for M sufficiently large

$$\begin{aligned}
\zeta &\leq \sum_{k=2CM^{\frac{d-1}{d}}}^{\infty} M^k M^{-\alpha k} M^{-(2^d-1)(k-CM^{\frac{d-1}{d}})c_1(1-\epsilon)} \\
&\leq \sum_{k=2CM^{\frac{d-1}{d}}}^{\infty} M^{(1-\alpha-(2^d-1)c_1(1-\epsilon))k} M^{(2^d-1)CM^{\frac{d-1}{d}}c_1(1-\epsilon)} \\
&= \frac{M^{(1-\alpha-(2^d-1)c_1(1-\epsilon))2CM^{\frac{d-1}{d}}}}{1 - M^{-1}} M^{(2^d-1)CM^{\frac{d-1}{d}}c_1(1-\epsilon)} \\
&= \frac{M}{M-1} M^{(1-\alpha-\frac{1}{2}(2^d-1)c_1(1-\epsilon))2CM^{\frac{d-1}{d}}} \leq M^{-\gamma},
\end{aligned}$$

where γ is arbitrarily large, provided α is chosen appropriately, and so $\zeta < 1/n$ for large enough n. □

References

[1] Andrew R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic Publishers, 1991.

[2] G. Blanchard and D. Geman. Hierarchical testing designs for pattern recognition. to appear in Annals of Statistics, 2005.

[3] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1983.

[4] M. V. Burnashev and K. Sh. Zigangirov. An interval estimation problem for controlled observations. *Problems in Information Transmission*, 10:223–231, 1974.

[5] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, pages 129–145, 1996.

[6] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[7] Cecil C. Craig. On the tchebychef inequality of bernstein. *The Annals of Statistics*, 4(2):94–102, 1933.

[8] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, 1st edition, 1990.

[9] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Information, prediction, and query by committee. *Proc. Advances in Neural Information Processing Systems*, 1993.

[10] P. Hall and I. Molchanov. Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *The Annals of Statistics*, 31(3):921–941, 2003.

[11] Jarvis Haupt and Robert Nowak. Signal reconstruction from noisy random projections. Technical report, University of Wisconsin, Madison, March 2005. Submitted to IEEE Transactions on Information Theory (available at http://www.ece.wisc.edu/∼nowak/infth.pdf).

[12] Alexander Korostelev. On minimax rates of convergence in image models under sequential design. *Statistics & Probability Letters*, 43:369–375, 1999.

[13] Alexander Korostelev and Jae-Chun Kim. Rates of convergence for the sup-norm risk in image models under sequential designs. *Statistics & probability Letters*, 46:391–399, 2000.

[14] A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction*. Springer Lecture Notes in Statistics, 1993.

[15] B. Laurent and P. Massard. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[16] D. J. C. Mackay. Information-based objective functions for active data selection. *Neural Computation*, 4:698–714, 1991.

[17] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

[18] K. Sung and P. Niyogi. Active learning for function approximation. *Proc. Advances in Neural Information Processing Systems*, 7, 1995.

[19] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Mathématiques et Applications, 41. Springer, 2004.

[20] R. Willett, A. Martin, and R. Nowak. Backcasting: Adaptive sampling for sensor networks. In *Proc. Information Processing in Sensor Networks*, 26-27 April, Berkeley, CA, USA, 2004.