Minimax Bounds for Active Learning

Rui M. Castro, Robert D. Nowak, *Senior Member, IEEE,*

Abstract

This paper analyzes the potential advantages and theoretical challenges of "active learning" algorithms. Active learning involves sequential sampling procedures that use information gleaned from previous samples in order to focus the sampling and accelerate the learning process relative to "passive learning" algorithms, which are based on non-adaptive (usually random) samples. There are a number of empirical and theoretical results suggesting that in certain situations active learning can be significantly more effective than passive learning. However, the fact that active learning algorithms are feedback systems makes their theoretical analysis very challenging. This paper aims to shed light on achievable limits in active learning. Using minimax analysis techniques, we study the achievable rates of classification error convergence for broad classes of distributions characterized by decision boundary regularity and noise conditions. The results clearly indicate the conditions under which one can expect significant gains through active learning. Furthermore we show that the learning rates derived are tight for "boundary fragment" classes in d-dimensional feature spaces when the feature marginal density is bounded from above and below.

Index Terms

Statistical Learning Theory, Minimax Lower Bounds, Active Learning, Adaptive Sampling

I. Introduction

Most theory and methods in machine learning focus on statistical inference based on a sample of independent and identically distributed (i.i.d.) observations. We call this typical set-up *passive learning* since the learning algorithms themselves have no influence in the data collection process. As widespread as the passive learning model is, in certain situations it is possible to combine the data collection and analysis processes, using data previously collected to guide in the selection of new samples. Sequential sampling strategies of this nature are called *active learning* procedures. Active learning can offer significant advantages over i.i.d. data collection.

To illustrate the idea of active learning, consider the prototypical example of document classification. Suppose we are given a text document and want to associate a topic/label to it (e.g., finance, sports, nuclear physics, information theory). Our goal is to devise an algorithm that learns how to perform this task from examples. That is, we have access to a number of documents that have been inspected and labeled by an expert (most likely a human), and

R. Castro is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, 53706 USA, e-mail: (see http://homepages.cae.wisc.edu/∼rcastro).

R. Nowak is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, 53706 USA, e-mail: (see http://ece.wisc.edu/∼nowak).

Supported by NSF grants CCR-0350213 and CNS-0519824.

the learning algorithm uses these instances to construct a general labeling rule for documents. In today's world of electronic media, we have a virtually infinite supply of documents at our fingertips. If the labeled documents in this scenario are i.i.d. (i.e., collected completely at random), then this corresponds to the usual passive learning model. However, labeling documents for training purposes is expensive and time consuming, and ideally we would like our algorithm to learn to correctly label documents based on a modest number of labeled examples. To accomplish this, we would like the algorithm to automatically select unlabeled documents that it has difficulty in labeling itself or whose label is potentially very informative, and then request the correct labels for these documents from an expert (human). The hope is that as the algorithm learns, it makes fewer and fewer requests for labels, and in this way the total number of labeled documents required for the learning task may be much smaller than needed if an arbitrary set of labeled documents were used instead.

Interest in active learning has increased greatly in recent years, in part due to the dramatic growth of data sets and the high cost of labeling examples in such sets. There are several empirical and theoretical results suggesting that in certain situations active learning can be significantly more efficient than passive learning [1], [2], [3], [4], [5]. Many of these results pertain to the "noiseless" setting, in which the labels are deterministic functions of the features (e.g., attributes extracted from documents such as the frequencies of keywords, etc.). In certain noiseless scenarios it has been shown that the number of labeled examples needed to achieve a desired classification error rate is much smaller than what would be needed using passive learning. In fact for some of those scenarios, active learning requires only $O(\log n)$ labeled examples to achieve the same performance that can be achieved through passive learning with n labeled examples [3], [6], [7], [8]. This exponential speed-up in learning rates is a tantalizing example of the power of active learning.

Although the noiseless setting is interesting from a theoretical perspective, it is very restrictive, and seldom relevant for practical applications. Some results have been obtained for active learning in the "bounded noise rate" setting. In this setting labels are no longer a deterministic function of the features, rather for a given feature the probability of one label is significantly higher than the probability of any other label. In the case of binary classification this means that if $(\boldsymbol{X}, Y)$ is a feature-label pair, where $Y \in \{0, 1\}$, then $|\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) - 1/2| \geq c$ for every $\boldsymbol{x}$ in the feature space, with $c > 0$. In other words, $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ "jumps" at the decision boundary, providing a very strong cue that active learning algorithms can use. In fact, this cue is effectively as strong as in the noiseless case. Under the bounded noise rate assumption it can be shown that results similar to the ones for the noiseless scenario can be achieved [4], [9], [10], [11]. These results are intimately related to coding with noiseless feedback [12], [13] and active sampling techniques in regression analysis [13], [14], [15], [10], [16], where related performance gains have been reported. Furthermore, the active learning algorithm proposed in [9], in addition to providing improvements in certain bounded noise conditions, is shown to perform no worse than passive learning in general conditions.

In this paper, we expand the theoretical investigation of active learning to include cases in which the noise is unbounded. In the case of binary classification this means that $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ is not bounded away from $1/2$. Notice that in this case there is no strong cue that active learning algorithms can follow, since the labels of

features near the decision boundary are almost devoid of information (i.e., $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ approaches $1/2$). Since situations like this seem very likely to arise in practice (e.g., simply due to feature measurement or precision errors if nothing else), it is important to identify the potential of active learning in such cases.

Our main result can be summarized as follows. Following Tsybakov's formulation of distributional classes [17], the complexity of classification problems can in many cases be characterized by two key parameters. The parameter $\rho = (d-1)/\alpha$, where d is the dimension of the feature space and α is the Hölder regularity of the Bayes decision boundary, is a measure of the complexity of the boundary. The parameter $\kappa \geq 1$ characterizes the level of "noise", that is, the behavior of $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ in the vicinity of the boundary. The value $\kappa = 1$ corresponds to the noiseless or bounded noise situation and $\kappa > 1$ corresponds to unbounded noise conditions. Using this sort of characterization, we derive lower and upper bounds for active learning performance. In particular, it is shown that the fastest rate of classification error decay using active learning is $n^{-\frac{\kappa}{2\kappa+\rho-2}}$, where n is the number of labeled examples, whereas the fastest decay rate possible using passive learning is $n^{-\frac{\kappa}{2\kappa+\rho-1}}$. Note that the active learning error decay rate is always faster than that of passive learning. Tsybakov has shown that in certain cases, when ($\kappa \to 1$ and $\rho \to 0$) passive learning can achieve "fast" rates approaching n^{-1} (faster than the usual $n^{-1/2}$ rate). In contrast, our results show that in similar situations active learning can achieve much faster rates (in the limit decaying as fast as any negative power of n). Also note that the passive and active rates are essentially the same as $\kappa \to \infty$, which is the case in which $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ is very flat near the boundary and consequently there is no cue that can effectively drive an active learning procedure. Furthermore, upper bounds show that the learning rates derived are tight for "boundary fragment" classes in d-dimensional feature spaces when the density of the marginal distribution $P_{\boldsymbol{X}}$ (over features) is bounded from above and below on $[0, 1]^d$.

Although extremely appealing, most practical active learning methods so far are plagued by many problems that prevent their application in realistic settings. In many settings they tend to perform very well when only a few labeled examples are provided, but as soon as that number increases their performance degrades significantly, often becoming worse than passive methods [18]. This is an indication that these learners are often "side-tracked" by the first few labeled examples. This behavior partially motivates the work presented here. By carefully characterizing the fundamental limits of active learning one hopes to be able to design sound practical algorithms not displaying the pitfalls of currently existing techniques.

There is a large literature on problems that bear some similarities to the investigation we are considering. One related area is known collectively as adaptive sampling (see [19] and references therein), dealing with inference problems where the sampling designs are adjusted during experiments. In contrast to our non-parametric focus here, the adaptive sampling literature addresses problems involving estimation or testing of scalar or parametric quantities (e.g., population sizes). Also somewhat related is the field of sequential analysis [20], [21], dealing with problems such as sequential hypothesis testing. However, as also pointed out in [14], the nature of those results is very different, since there is no adjustment of the sampling procedure during the inference process.

The paper is organized as follows. In Section II we formally state the active learning problem and the basic questions we are interested in. Section III addresses the performance of active learning for one-dimensional threshold

classifiers, under a characterization of the noise condition. Upper and lower bounds for the active learning rates are presented. In Section IV we extend these results to the multidimensional class of boundary fragments, deriving performance lower bounds and corresponding upper bounds, and demonstrating that these results are near minimax optimal. Final remarks are made in Section V and the main proofs are given in the Appendices.

II. Problem Formulation

Let $\mathcal{X} \stackrel{\Delta}{=} [0,1]^d$ denote the *feature space* and $\mathcal{Y} \stackrel{\Delta}{=} \{0,1\}$ denote the *label space*. Let $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random vector, with *unknown* distribution $P_{\boldsymbol{X}Y}$. The goal in classification is to construct a "good" classification rule, that is, given a feature vector $\boldsymbol{X} \in \mathcal{X}$ we want to predict the label $Y \in \mathcal{Y}$ as accurately as possible, where the classification rule is a measurable function $f : \mathcal{X} \to \mathcal{Y}$. The performance of the classifier is evaluated in terms of the expected 0/1-loss. With this choice of loss function the risk is simply the probability of classification error,

$$R(f) \stackrel{\Delta}{=} \mathbb{E}[\mathbf{1}\{f(\boldsymbol{X}) \neq Y\}] = \Pr(f(\boldsymbol{X}) \neq Y) \ ,$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Since we are considering only binary classification (two classes) there is a one-to-one correspondence between classifiers and sets: Any reasonable deterministic classifier is of the form $f(\boldsymbol{x}) = \mathbf{1}\{\boldsymbol{x} \in G\}$, where G is a measurable subset of $[0,1]^d$. We use the term classifier interchangeably for both f and G. Define the optimal risk as

$$R^* \stackrel{\Delta}{=} \inf_{G \text{ measurable}} R(G) \ .$$

A classifier attaining the minimal risk R^* is the *Bayes Classifier* $G^* \stackrel{\Delta}{=} \{\boldsymbol{x} \in [0,1]^d : \eta(\boldsymbol{x}) \geq 1/2\}$, where

$$\eta(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \ ,$$

is called the *feature conditional probability*. We use the term conditional probability only if it is clear from the context. In general $R(G^*) = R^* > 0$ unless the labels are a deterministic function of the features, and therefore even the optimal classifier misclassifies sometimes. For that reason the quantity of interest for the performance evaluation of a classifier G is the *excess risk* (or *regret*)

$$R(G) - R(G^*) = \int_{G \Delta G^*} |2\eta(\boldsymbol{x}) - 1| \mathrm{d}P_{\boldsymbol{X}}(\boldsymbol{x}) \ , \tag{1}$$

where Δ denotes the symmetric difference between two sets[1], and $P_{\boldsymbol{X}}$ is the marginal distribution of $\boldsymbol{X}$.

As mentioned before $P_{\boldsymbol{X}Y}$ is generally unknown, therefore direct construction of the Bayes classifier is impossible. Suppose that we have available a large (infinite) pool of feature examples we can choose from, large enough so that we can choose any feature point $\boldsymbol{X}_i \in [0,1]^d$ and observe its label Y_i. The data collection has a temporal aspect to it, namely we collect the labeled examples one at the time, starting with $(\boldsymbol{X}_1, Y_1)$ and proceeding until $(\boldsymbol{X}_n, Y_n)$ is observed. One can view this process as a query learning procedure, in which one queries the label of a feature vector. Formally we have:

[1]Define $A \Delta B \stackrel{\Delta}{=} (A \cap B^c) \cup (A^c \cap B)$, where A^c and B^c are the complement of A and B respectively.

A1 - Given the feature vector $\boldsymbol{X}_i$, $i \in \{1, \ldots, n\}$, the label $Y_i \in \{0, 1\}$, is such that

$$\Pr(Y_i = 1 | \boldsymbol{X}_i) = \eta(\boldsymbol{X}_i) \ .$$

The random variables $\{Y_i\}_{i=1}^n$ are conditionally independent given $\{\boldsymbol{X}_i\}_{i=1}^n$.

A2.1 - Passive Sampling: $\boldsymbol{X}_i$ is independent of $\{Y_j\}_{j \neq i}$.

A2.2 - Active Sampling: $\boldsymbol{X}_i$ depends only on $\{\boldsymbol{X}_j, Y_j\}_{j<i}$. In other words $\{\boldsymbol{X}_j\}$ obeys a causal relation of the form

$$\boldsymbol{X}_i = h_i(\boldsymbol{X}_1 \ldots \boldsymbol{X}_{i-1}, Y_1 \ldots Y_{i-1}, \boldsymbol{U}_i) \ ,$$

where $h_i(\cdot)$ is a deterministic function, and $\boldsymbol{U}_i$ accounts for possible randomization of the sampling rule. In other words $\boldsymbol{U}_i$ is a random variable, independent of $(\boldsymbol{X}_1 \ldots \boldsymbol{X}_{i-1}, Y_1 \ldots Y_{i-1})$.

The function $h_i(\cdot)$, together with $\boldsymbol{U}_i$, is called the the sampling strategy at time i. The collection of sampling strategies for all $i \in \{1, \ldots, n\}$ is called the *sampling strategy*, denoted by S_n. It completely defines the sampling procedure. After collecting the n labeled examples, that is after collecting $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$, we construct a classifier $\widehat{G}_n$ that is desirably "close" to G^* in terms of excess risk. The subscript n denotes dependence on the data set, instead of writing it explicitly.

Under the passive sampling scenario (A2.1) the sample locations do not depend on the labels (except for the trivial dependence between $\boldsymbol{X}_i$ and Y_i), and therefore the collection of sample points $\{\boldsymbol{X}_i\}_{i=1}^n$ can be chosen before any observations (i.e., labels) are collected. On the other hand the active sampling scenario (A2.2) allows for the i^{th} sample location to be chosen using all the information collected up to that point (the previous $i-1$ samples). It is clear that (A2.2) is more general than (A2.1), with the former assumption allowing for much more flexibility. It is important to remark that in the active sampling setting (A2.2) the learning algorithm is actively choosing the feature $\boldsymbol{X}_i$. This is often referred to as *adaptive sampling* or *query learning*, where the learner can make queries to an expert, requesting labels for features from synthetic examples [22]. We avoid the use of the term "adaptive" in this paper since it is often used to refer to adaptivity of the inference strategy to unknown regularity parameters. We use the term "active sampling" instead, to prevent confusion. Other active learning paradigms exist, for example *pool-based learning*, where we assume a large (infinite) pool of feature examples for which we can ask for a label. If we assume an infinite (or nearly infinite) pool of unlabeled examples, and that the marginal feature density $p_{\boldsymbol{X}}$ is bounded away from zero, then we can essentially choose any feature point $\boldsymbol{X}_i$ for labeling, since the set of examples in that pool is dense in $\mathcal{X}$, because $p_{\boldsymbol{X}}$ is bounded from below.

Some clarification pertaining the role of $P_{\boldsymbol{X}}$, the marginal distribution of $\boldsymbol{X}$, is important at this point. Recall that this marginal distribution partially dictates how the error is measured (1). The passive and active sampling paradigms, respectively (A2.1) and (A2.2), allow the collection of labels corresponding to arbitrary feature vectors, and so the sampling processes do not provide information about $P_{\boldsymbol{X}}$. Therefore, in our error calculations we take $P_{\boldsymbol{X}}$ to be uniform, which suffices for the purposes of determining rates of convergence for all cases in which $P_{\boldsymbol{X}}$ is

bounded from above and away from zero (since under those assumptions controlling $\int_{G\Delta G^*} |2\eta(\boldsymbol{x}) - 1| \mathrm{d}\boldsymbol{x}$ suffices to control the excess risk $R(G) - R(G^*)$). There are other active learning paradigms that may not require such assumptions, namely *selective sampling* (e.g., see [3], [4]), where at each time step the algorithm is presented a feature vector drawn independently from $P_{\boldsymbol{X}}$ and must decide whether or not to collect the respective label.

To be able to present performance guarantees on the excess risk behavior we need to impose further conditions on the possible distributions $P_{\boldsymbol{X}Y}$. We are particularly interested in the framework proposed by Tsybakov in [17], consisting of a characterization of the regularity of the Bayes decision sets, and the behavior of the conditional probability η in the vicinity of the Bayes decision boundary.

III. One-Dimensional Threshold Classifiers ($d = 1$)

In this section we consider a class of problems in which the Bayes classifier is a threshold function. Although this corresponds to a rather simple class of distributions, a complete characterization of achievable performance for active learning in this class was previously unknown. Moreover, the study of this simple class sheds light on the potential advantages and limitations of active learning, and provides crucial understanding to tackle more complicated problems. Throughout this section the feature space is the unit interval $[0, 1]$. Let P_{XY} be the distribution governing $(X, Y) \in [0, 1] \times \{0, 1\}$. Assume that the Bayes classifier for this distribution is of the form $[\theta^*, 1]$, which means that $\eta(x) < 1/2$ for all $x < \theta^*$ and $\eta(x) \geq 1/2$ for all $x \geq \theta^*$. We assume that p_X, the marginal density of X with respect to the Lebesgue measure, is uniform on $[0, 1]$, although the results in this paper are easily generalized to the case where that marginal density is not uniform, but bounded above and below (in which case one obtains exactly the same rates of excess risk convergence). In order to gain a deeper understanding of the potential of active learning we impose further conditions on η, characterizing the behavior of the conditional probability around the Bayes decision boundary. For $\kappa \geq 1$ and $c > 0$ we assume that

$$|\eta(x) - 1/2| \geq c|x - \theta^*|^{\kappa - 1} \ , \tag{2}$$

for all x such that $|\eta(x) - 1/2| \leq \delta$, with $\delta > 0$ (with $0^0 \triangleq \lim_{t \to 0^+} t^0 = 1$).

The condition above is similar to the "noise-condition" introduced by Tsybakov [17]. In fact it is easily shown that distributions satisfying the conditions above (including (2)) also satisfy the noise condition stated in [17]. Condition (2) indicates that $\eta(\cdot)$ cannot be arbitrarily "flat" around the decision boundary and plays a critical role on the performance of any classification rule obtained through labeled examples. We also assume a reverse-sided condition on $\eta(\cdot)$, namely

$$|\eta(x) - 1/2| \leq C|x - \theta^*|^{\kappa - 1} \ , \tag{3}$$

for all $x \in [0, 1]$, where $C > c$. This condition, together with (2), provides a two-sided characterization of the "noise" around the decision boundary. Similar two-sided conditions have been proposed for other problems [23], [24]. The condition above can be made local, where (3) holds only for x in the vicinity of θ^*. However, local conditions of this type do not give rise to larger distribution classes, and therefore we do not consider such generalizations.

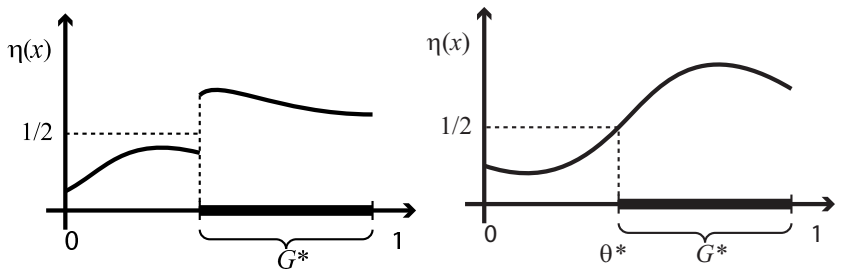


Fig. 1. *Examples of two conditional distributions $\eta(x) = \Pr(Y = 1|X = x)$. (a) In this case $\eta(\cdot)$ satisfies the margin condition with $\kappa = 1$; (b) Here the margin condition is satisfied for $\kappa = 2$.*

Let $\mathcal{P}(\kappa, c, C)$ be the class of distributions with uniform marginal P_X and satisfying (2) and (3). If $\kappa = 1$ then the $\eta(\cdot)$ function "jumps" across 1/2, that is $\eta(\cdot)$ is bounded away from the value $1/2$ (see Figure 1(a)). If $\kappa > 1$ then $\eta(\cdot)$ crosses the value $1/2$ at θ^*. An especially interesting case (from a practical perspective) corresponds to $\kappa = 2$ (Figure 1(b)), where the conditional probability behaves linearly around $1/2$. This situation is perhaps most likely to arise in practice for the following reason. If the observed features are contaminated with noise (e.g., i.i.d. additive errors), then $\eta(\cdot)$ is effectively smoothed, and thus typically everywhere differentiable with non-vanishing first derivative near the Bayes decision boundary (equivalently $\kappa = 2$). Finally if $\kappa > 2$, then the first derivative of $\eta(\cdot)$ is zero and $\eta(\cdot)$ is "flat" around θ^*.

We begin by presenting lower bounds on the performance of active and passive learning methods for the class of distributions $\mathcal{P}(\kappa, c, C)$. Most of the results that follow involve multiplicative *constant factors*, that is factors that do not depend on the sample size n. We denote these by the symbol const, generally without explicitly describing them.

Theorem 1 (Active Learning Minimax Lower Bound for $d = 1$)**.** *Let $\kappa > 1$. Under the assumptions (A1) and (A2.2) we have*

$$\inf_{\widehat{G}_n, S_n} \sup_{P_{XY} \in \mathcal{P}(\kappa,c,C)} \mathbb{E}\left[R(\widehat{G}_n) - R(G^*)\right] \geq \mathrm{const}(\kappa, c, C)\ n^{-\frac{\kappa}{2\kappa-2}}\ ,$$

for n large enough, where $\mathrm{const}(\kappa, c, C) > 0$ and the infimum *is taken over the set of all possible classification rules $\widehat{G}_n$ and active sampling strategies S_n. Note also that condition* (3) *does not play a prominent role in the result, in other words even when $C = \infty$ we have $\mathrm{const}(\kappa, c, C) < \infty$.*

The theorem is proved in Appendix A. The proof employs relatively standard techniques, and follows the approach in [25]. The key idea is to reduce the original problem to the problem of deciding among a finite collection of representative distributions. The determination of an appropriate collection of such distributions and careful management of assumption (A2.2) are the key aspects of the proof.

Contrast this result with the one attained for passive learning. Under the passive learning model it is clear that the sample locations $\{X_i\}_{i=1}^n$ must be scattered around the interval $[0, 1]$ in a somewhat uniform manner. These

can be deterministically placed, for example over a uniform grid, or simply taken uniformly distributed over $[0,1]$. Using similar lower bounding techniques we obtain be following result.

Proposition 1 (Passive Learning Minimax Lower Bound for $d=1$)**.** *Under assumptions (A1), (A2.1), and $\kappa \geq 1$, we have*

$$\inf_{\widehat{G}_n} \sup_{P_{XY} \in \mathcal{P}(\kappa,c,C)} \mathbb{E}\left[R(\widehat{G}_n) - R(G^*)\right] \geq \text{const}(\kappa,c,C)\ n^{-\frac{\kappa}{2\kappa-1}}\ , \tag{4}$$

where the samples $\{X_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) uniformly over $[0,1]$.

The proof, which is a slight modification of the proof of Theorem 1, is sketched in Appendix B. Furthermore the bound is tight, in the sense that it is possible to devise classification strategies attaining the same asymptotic excess risk behavior [17].

We notice that under the passive learning model the excess risk decays at a strictly slower rate than for the active sampling scenario. The difference is dramatic as $\kappa \to 1$. If $\kappa = 1$ (Figure 1(a)) it can actually be shown that an exponential rate of error decay is attainable by active sampling [13]. As $\kappa \to \infty$ the excess risk decay rates become similar, regardless of the sampling paradigm (either (A2.1) or (A2.2)). This indicates that if no assumptions are made on the conditional distribution $\Pr(Y=1|X=x)$ then no advantage can be gained from the extra flexibility of active sampling. The intuition is that active learning can only help (in a minimax sense) if there is a strong enough cue indicating the decision boundary. If $\eta(\cdot)$ is arbitrarily "flat" near the boundary, then there is essentially no such cue. Assumption 2 plays a critical role quantifying the strength of the cue.

As remarked before a most relevant case is $\kappa = 2$. In this case, the rate for active learning is n^{-1}, which is significantly faster than $n^{-2/3}$, the best possible rate for passive learning. Also observe that the difference between active and passive learning rates becomes arbitrarily large as $\kappa \to 1$, with the excess risk of active learning tending to decay faster than n^{-p} for any $p > 0$, and that of passive learning tending to decay like n^{-1}.

Next we describe a methodology showing that the rates of Theorem 1 are nearly achievable. We start by presenting an algorithm proposed by Burnashev and Zigangirov [13], inspired by an idea of Horstein [12]. This algorithm is designed to work in the bounded noise case, that is when $\kappa = 1$, corresponding to a scenario where the conditional probability $\eta(x) = \Pr(Y=1|X=x)$ is bounded away from $1/2$, that is $|\eta(x) - 1/2| \geq c$ for all $x \in [0,1]$. This algorithm is then adapted to handle situations in which $\kappa > 1$.

A. Bounded noise rate ($\kappa = 1$)

First let us suppose that the observations are noiseless (that is $\eta(x) \in \{0,1\}$). Then it is clear that one can estimate the Bayes decision boundary θ^* very efficiently using binary bisection: start by taking a sample at $X_1 = 1/2$. Depending on the resulting label $Y_1|X_1$ we know if θ^* is to the left of X_1 (if $Y_1 = 1$) or to the right (if $Y_1 = 0$). Proceeding accordingly we can construct an estimate of θ^* denoted by $\widehat{\theta}_n$ and a corresponding classifier $\widehat{G}_n \triangleq [\widehat{\theta}_n, 1]$ such that

$$R(\widehat{G}_n) - R(G^*) = |\widehat{\theta}_n - \theta^*| \leq 2^{-(n+1)}\ .$$

Now let us assume that some level of noise is present, but that the bounded noise condition, $|\eta(x) - 1/2| \geq c$ for all $x \in [0,1]$, is met (note that $0 \leq c \leq 1/2$). The learning task is more complicated in this case because deciding where to sample depends on noisy and therefore somewhat unreliable label observations. Nevertheless there is a probabilistic bisection method, proposed in [12], that is suitable for this purpose. The key idea stems from Bayesian estimation. Suppose that we have a prior probability density function $p_0(\cdot)$ on the unknown parameter θ^*, namely that θ^* is uniformly distributed over the interval $[0,1]$ (that is $p_0(\theta^* = x) = \mathbf{1}\{x \in [0,1]\}$). To illustrate the process, let us assume the particular scenario that $\theta^* = 1/4$. As in the noiseless case, begin by taking a measurement at $X_1 = 1/2$, that is collect Y_1 given X_1. With probability at least $\eta(X_1) \geq 1/2 + c$ we observe a one, and with probability at most $1 - \eta(X_1) \leq 1/2 - c$ we observe a zero. Therefore it is more likely to observe a one than a zero. Assume for example that $Y_1 = 1$. Given these facts we can compute a "posterior" density $p_1(\cdot)$ simply by applying an approximate Bayes rule (for the application of Bayes rule we assume the most difficult/noisy setting, $|\eta(X_1) - 1/2| = c$). In this case we get

$$p_1(\theta^* = x | X_1 = 1/2, Y_1 = 1) = \begin{cases} 1 + 2c & \text{, if } x \leq 1/2, \\ 1 - 2c & \text{, if } x > 1/2, \end{cases} .$$

The next step is to choose the sample location X_2. We choose X_2 so that is *bisects* the posterior distribution, that is, we take X_2 such that $\Pr_{\theta \sim p_1}(\theta > X_2 | X_1, Y_1) = \Pr_{\theta \sim p_1}(\theta < X_2 | X_1, Y_1)$. In other words X_2 is just the median of the posterior distribution. If our model is correct, the probability of the event $\{\theta < X_2\}$ is identical to the probability of the event $\{\theta > X_2\}$ and therefore sampling at X_2 seems to be most informative. We continue iterating this procedure until we have collected n samples. The estimate $\widehat{\theta}_n$ is defined as the median of the final posterior distribution. Note that if $c = 1/2$ then probabilistic bisection is simply the binary bisection described above.

The above algorithm seems to work extremely well in practice, but it is difficult to analyze and we are not aware of theoretical guarantees for it, especially pertaining rates of error decay. In [13] an algorithm inspired by that approach was presented. Although its operation is slightly more complicated, it is easier to analyze. The proposed algorithm uses essentially the same ideas but enforces a discrete structure for the posterior, by forcing the samples X_i to lie on a grid, in particular $X_i \in \{0, t, 2t, \ldots, 1\}$ where $m = t^{-1} \in \mathbb{N}$. A description of the algorithm can be found in [13] or in [26]. We call this method the Burnashev-Zigangirov (BZ) algorithm. This algorithm returns a confidence interval $\widehat{I}_n = [t(i-1), ti]$, $i \in \{1, \ldots, m\}$, such that with high probability $\theta^* \in \widehat{I}_n$. In fact we have the following remarkable result [13]:

$$\Pr(\theta^* \notin \widehat{I}_n) \leq \frac{1}{t}\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4c^2}\right)^n \leq \frac{1}{t}\left(1 - c^2\right)^n \leq \frac{1}{t}\exp(-nc^2) , \tag{5}$$

where the last two steps follow from the fact that $\sqrt{x} \leq (x+1)/2$ for all $x \geq 0$, and that $(1+s)^x \leq \exp(xs)$ for all $x > 0$ and $s > -1$. An estimate $\widehat{\theta}_n$ can be constructed using the mid-point of the interval $\widehat{I}_n$ (note that $\widehat{I}_n$ has length t), and a candidate classification set/rule is $\widehat{G}_n = [\widehat{\theta}_n, 1]$. To get a bound on the expected excess risk one

proceeds using the standard integration approach.

$$\begin{aligned}
\mathbb{E}[R(\widehat{G}_n)] - R(G^*) &= \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} |2\eta(x) - 1| \mathrm{d}x\right] \\
&\leq \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} \mathrm{d}x\right] \\
&= \mathbb{E}\left[|\widehat{\theta}_n - \theta^*|\right] \\
&= \int_0^1 \Pr\left(|\widehat{\theta}_n - \theta^*| > z\right) \mathrm{d}z \\
&= \int_0^{t/2} \Pr\left(|\widehat{\theta}_n - \theta^*| > z\right) \mathrm{d}z + \int_{t/2}^1 \Pr\left(|\widehat{\theta}_n - \theta^*| > z\right) \mathrm{d}z \\
&\leq t/2 + (1 - t/2) \Pr\left(|\widehat{\theta}_n - \theta^*| > t/2\right) \\
&\leq t/2 + \frac{1}{t} \exp(-nc^2) \ .
\end{aligned}$$

Taking $t = \sqrt{2}\exp(-nc^2/2)$ yields

$$\mathbb{E}[R(\widehat{G}_n)] - R(G^*) \leq \sqrt{2}\exp(-nc^2/2) \ .$$

Notice that the excess risk decays exponentially in the number of samples. This is much faster than what is attainable using passive sampling, where the decay rate is $1/n$. This is the same error behavior as in the noiseless scenario, where we have an exponential rate of error decay using binary bisection. The difference is that now the exponent depends on the noise margin c, larger noise margins corresponding to faster error decay rates. In [13] some guarantees on the faster exponential decay rate are also given, as a function of c.

B. Unbounded rate noise: $\kappa > 1$

In this section we consider scenarios where the noise rate is not bounded, that is, observations made closer to the transition point θ^* are noisier than observations made further away. In light of Theorem 1 this degradation of observation quality hinders extremely fast excess risk decay rates.

To study this case, we proceed as in the case $\kappa = 1$. Collect samples over a grid, namely $X_i \in \{0, t, 2t, \ldots, 1\}$ where $m = t^{-1} \in \mathbb{N}$. Assume for a brief moment that the grid is not aligned with the transition point θ^*, for example $|\theta^* - kt| \geq t/3$ for all $k \in \{0, \ldots, m\}$. This implies that $|\eta(x) - 1/2| \geq c(t/3)^{\kappa-1}$ for all $x \in \{0, t, \ldots, 1\}$ (assume that t is small enough so that $\delta \geq c(t/3)^{\kappa-1}$). Of course this is in general an unrealistic assumption, since θ^* may be arbitrarily close to a grid point, but let us assume this condition for now. We can proceed by using the algorithm described in the previous section replacing c by $c(t/3)^{\kappa-1}$ and using (5). Notice also that due to (3) the behavior

of the expected excess risk is related to the behavior of $|\widehat{\theta}_n - \theta^*|$ in an interesting way,

$$\begin{aligned}
\mathbb{E}[R(\widehat{G}_n)] - R(G^*) &= \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} |2\eta(x) - 1| \mathrm{d}x\right] \\
&\leq 2C\ \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} |x - \theta^*|^{\kappa-1} \mathrm{d}x\right] \\
&= 2C\ \mathbb{E}\left[\int_{\min\{\theta^*, \widehat{\theta}_n\}}^{\max\{\theta^*, \widehat{\theta}_n\}} |x - \theta^*|^{\kappa-1} \mathrm{d}x\right] \\
&= \frac{2C}{\kappa} \mathbb{E}[|\widehat{\theta}_n - \theta^*|^{\kappa}]\ .
\end{aligned}$$

We now proceed in a similar fashion as before

$$\begin{aligned}
\mathbb{E}[R(\widehat{G}_n)] - R(G^*) \leq \frac{2C}{\kappa} \mathbb{E}\left[|\widehat{\theta}_n - \theta^*|^{\kappa}\right] &= \frac{2C}{\kappa} \int_0^1 \Pr\left(|\widehat{\theta}_n - \theta^*|^{\kappa} > z\right) \mathrm{d}z \\
&= \frac{2C}{\kappa} \int_0^1 \Pr\left(|\widehat{\theta}_n - \theta^*| > z^{1/\kappa}\right) \mathrm{d}z \\
&= \frac{2C}{\kappa} \left[\int_0^{(t/2)^{\kappa}} \Pr\left(|\widehat{\theta}_n - \theta^*| > z^{1/\kappa}\right) \mathrm{d}z \right. \\
&\quad \left. + \int_{(t/2)^{\kappa}}^1 \Pr\left(|\widehat{\theta}_n - \theta^*| > z^{1/\kappa}\right) \mathrm{d}z\right] \\
&\leq \frac{2C}{\kappa} \left[(t/2)^{\kappa} + (1 - (t/2)^{\kappa}) \Pr\left(|\widehat{\theta}_n - \theta^*| > t/2\right)\right] \\
&\leq \frac{2C}{\kappa} \left[(t/2)^{\kappa} + \frac{1}{t} \exp\left(-nc^2 (t/3)^{2\kappa-2}\right)\right]\ ,
\end{aligned}$$

Finally, let

$$t = 3\left(\frac{\kappa+1}{c^2(2\kappa-2)}\ \frac{\log n}{n}\right)^{\frac{1}{2\kappa-2}}\ ,$$

to conclude that

$$\mathbb{E}[R(\widehat{G}_n) - R(G^*)] \leq \mathrm{const}(\kappa, c, C) \cdot \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}\ , \tag{6}$$

where $\mathrm{const}(\kappa, c, C) > 0$ is a constant factor. The error decay rate of this upper bound corresponds to the rate of lower bound in Theorem 1, apart from logarithmic factors. The result indicates that, in principle, a methodology similar to the BZ algorithm might allow us to achieve the lower bound rates.

It is important to emphasize that the above result holds under the assumption that the sampling grid is not aligned with the unknown threshold point θ^*. If this is not the case then we have $|\eta(x) - 1/2| < c(t/3)^{\kappa-1}$ for one of the sampling points, and the analysis above is no longer valid. This set-back can be avoided in different ways, for example using a direct modification of the BZ sampling strategy, as in [26], or using several sampling grids simultaneously. We describe the latter approach in what follows. Begin by dividing the available measurements into three sets of the same size, and use three sampling grids, each shifted slightly from each other (a precise description is given below). Suppose we have a budget of n samples (without loss of generality assume that n is

divisible by 3). We divide the budget in three, using $n/3$ samples and applying the BZ algorithm to each of the three sampling grids. The rational is that at most one of these sampling grids is going to be closely aligned with the unknown transition point θ^*, and therefore the other two cannot be aligned with θ^*. Thus, for at least two of the three resultant estimators we have provable performance bounds, and so it suffices to check for agreement among these three estimates: if at least two of them agree on the location of θ^*, then we take this majority decision as our final estimator $\widehat{\theta}_n$. Next we describe this procedure formally.

Consider three different sampling grids, $\text{Grid}^{(A)} = \{t, 2t, \ldots, 1-t\}$, $\text{Grid}^{(B)} = \{t/3, 4t/3, \ldots, 1-2t/3\}$, and $\text{Grid}^{(C)} = \{2t/3, 5t/3, \ldots, 1-t/3\}$. Samples can also be "taken" at $X_i \in \{0,1\}$ but in that case we impose that $Y_i = X_i$. This does not effect the algorithm performance. For at least two of the estimators we have $|\theta^* - x| \geq t/6$ for all the points in the respective sampling grids. This means that for samples x taken on those grids we have $|\eta(x) - 1/2| \geq c(t/6)^{\kappa-1}$ and therefore we can use the bounded noise rate results in the analysis. We now run the BZ algorithm three times, using $n/3$ samples each time, taken on the grids $\text{Grid}^{(A)}$, $\text{Grid}^{(B)}$ and $\text{Grid}^{(C)}$. Let $\widehat{I}_n^{(A)}$, $\widehat{I}_n^{(B)}$ and $\widehat{I}_n^{(C)}$ denote the confidence intervals returned by the three applications of the BZ algorithm. Next aggregate these confidence intervals to construct a final confidence interval $\widehat{I}_n$ as follows:

$$
\begin{array}{l}
\texttt{If } |\widehat{I}_n^{(A)} \cap \widehat{I}_n^{(B)}| = 2t/3 \texttt{ let } \widehat{I}_n = \widehat{I}_n^{(A)} \cup \widehat{I}_n^{(B)} \\
\quad \texttt{else if } |\widehat{I}_n^{(A)} \cap \widehat{I}_n^{(C)}| = 2t/3 \texttt{ then } \widehat{I}_n = \widehat{I}_n^{(A)} \cup \widehat{I}_n^{(C)} \\
\quad\quad \texttt{else if } |\widehat{I}_n^{(B)} \cap \widehat{I}_n^{(C)}| = 2t/3 \texttt{ then } \widehat{I}_n = \widehat{I}_n^{(B)} \cup \widehat{I}_n^{(C)} \\
\quad\quad\quad \texttt{else set } \widehat{I}_n = \{1/2\}
\end{array}
$$

Note that if at least two of the confidence intervals overlap (resulting in an intersection of length $2/3t$), then the final confidence interval length is $4/3t$. If none of the intervals overlap, then the default location is arbitrarily chosen to be $1/2$.

Proposition 2. *Proceeding as described above and assuming* (2) *we have*

$$\Pr(\theta^* \notin \widehat{I}_n) \leq \frac{2}{t} \exp\left(-\frac{n}{3} c^2 (t/6)^{2\kappa-2}\right) . \tag{7}$$

Proof: We need to consider two separate scenarios: either θ^* is "close" to a point in one of the sampling grids, or θ^* is close to zero or one. Consider the first situation and without loss of generality assume that θ^* is close to a point in $\text{Grid}^{(A)}$, that is there exists $x \in \text{Grid}^{(A)}$ such that $|\theta^* - x| < t/6$. This implies that for all $x \in \text{Grid}^{(B)} \cup \text{Grid}^{(C)}$ we have $|\theta^* - x| \geq t/6$, and therefore

$$\Pr(\theta^* \notin \widehat{I}_n^{(B)}) \leq \frac{1}{t} \exp\left(-\frac{n}{3} c^2 (t/6)^{2\kappa-2}\right) ,$$

and

$$\Pr(\theta^* \notin \widehat{I}_n^{(C)}) \leq \frac{1}{t} \exp\left(-\frac{n}{3} c^2 (t/6)^{2\kappa-2}\right) .$$

Define the event $E = \{\theta^* \in \widehat{I}_n^{(B)} \cap \widehat{I}_n^{(C)}\}$ *and notice that*

$$\Pr(E) \geq 1 - \frac{2}{t} \exp\left(-\frac{n}{3} c^2 (t/6)^{2\kappa-2}\right) \ .$$

Assume that E *holds. If* $|\widehat{I}_n^{(A)} \cap \widehat{I}_n^{(B)}| = 2t/3$ *then* $\theta^* \in \widehat{I}_n^{(B)} \subseteq \widehat{I}_n$*, if* $|\widehat{I}_n^{(A)} \cap \widehat{I}_n^{(C)}| = 2t/3$ *then* $\theta^* \in \widehat{I}_n^{(C)} \subseteq \widehat{I}_n$ *and, if* $|\widehat{I}_n^{(B)} \cap \widehat{I}_n^{(C)}| = 2t/3$ *then* $\theta^* \in \widehat{I}_n^{(B)} \subseteq \widehat{I}_n$*. Finally since there is a point in Grid*$^{(A)}$ *close to* θ^* *this implies that, under* E*, we have* $|\widehat{I}_n^{(B)} \cap \widehat{I}_n^{(C)}| = 2t/3$ *and so* $\widehat{I}_n \neq \{1/2\}$*. In conclusion, if* θ^* *is close to a point in Grid*$^{(A)}$ *then* $\Pr(\theta^* \notin \widehat{I}_n) \leq 1 - \Pr(E)$*, implying* (7)*.*

Now consider the case when θ^* *is close to zero or one. Without loss of generality suppose* θ^* *is close to zero, that is* $\theta^* < t/6$*. Then we have again that*

$$\Pr(\theta^* \in \widehat{I}_n^{(B)} \cap \widehat{I}_n^{(C)}) \geq 1 - \frac{2}{t} \exp\left(-\frac{n}{3} c^2 (t/6)^{2\kappa-2}\right) \ ,$$

and so a similar reasoning as above can be applied yielding the desired result. As a final remark note that when θ^* *is close to zero or one, the choice of intervals* $I_n^{(B)}$ *and* $I_n^{(C)}$ *above was arbitrary, and any of the three possible choices of pairs of intervals would work.* ■

Now we are in a similar situation as before, and taking $\widehat{\theta}_n$ as the midpoint of $\widehat{I}_n$ yields the bound

$$\Pr(|\widehat{\theta}_n - \theta^*| \geq \frac{2t}{3}) \leq \frac{2}{t} \exp\left(-\frac{n}{3} c^2 (t/6)^{2\kappa-2}\right) \ , \tag{8}$$

without requiring any assumptions on the location of θ^*. Simply proceeding as before shows that this algorithm provably attains the rate in (6).

Although the above methodology is satisfying from a theoretical point of view, it is somewhat wasteful (essentially only one third of the samples are effectively used), and does not generalize well to more complicated and realistic scenarios than the one considered in this paper. It is worth pointing out that in practice the basic BZ algorithm, without the modifications above, appears to work very well, even when the sampling grid is aligned with θ^*. The difficulties arise solely on the performance analysis. We conclude this section by summarizing the results in the following theorem.

Theorem 2. *Under assumptions* (2) *and* (3) *the active learning algorithms above satisfy*

$$\mathbb{E}[R(\widehat{G}_n) - R(G^*)] \leq \begin{cases} \sqrt{2} \exp(-nc^2/2) & \text{, if } \kappa = 1 \\ \text{const}(\kappa, c, C) \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}} & \text{, if } \kappa > 1 \end{cases} \ ,$$

for n *large enough, where* $\text{const}(\kappa, c, C) > 0$ *is a constant factor.*

Note that for $\kappa > 1$ the rate of error decay is almost as good as the rate in Theorem 1. This means that the methodology developed is nearly minimax optimal, since the rate in Theorem 2 differs only by a logarithmic factor in n. In the next section we generalize these results to a non-parametric setting in multiple dimensions.

IV. BOUNDARY FRAGMENTS ($d > 1$)

In this section we consider a much more general class of distributions, namely scenarios where the Bayes decision set is a boundary fragment. In other words the Bayes decision set is the epigraph of a function. We consider Hölder smooth boundary functions. Throughout this section we assume that $d \geq 2$, where d is the dimension of the feature space $[0,1]^d$.

Definition 1. *A function $f : [0,1]^{d-1} \to \mathbb{R}$ is* Hölder smooth*, with smoothness parameter $\alpha \geq 1$, if it has continuous partial derivatives up to order $k = \lfloor \alpha \rfloor$ (k is the maximal integer such that $k < \alpha$) and*

$$\forall\ \boldsymbol{z}, \boldsymbol{x} \in [0,1]^{d-1} : \quad |f(\boldsymbol{z}) - TP_{\boldsymbol{x}}(\boldsymbol{z})| \leq L\|\boldsymbol{z} - \boldsymbol{x}\|^{\alpha}\ ,$$

where $L > 0$ and $TP_{\boldsymbol{x}}(\cdot)$ denotes the degree k Taylor polynomial approximation of f expanded around $\boldsymbol{x}$. Denote this class of functions by $\Sigma(L, \alpha)$.

For any $g \in \Sigma(L,\alpha)$ let $\mathrm{epi}(g) \stackrel{\Delta}{=} \left\{(\boldsymbol{x}, y) \in [0,1]^{d-1} \times [0,1] : y \geq g(\boldsymbol{x})\right\}$ be the epigraph of g. Define

$$\mathcal{G}_{\mathrm{BF}} \ \stackrel{\Delta}{=}\ \{\mathrm{epi}(g) : g \in \Sigma(L,\alpha)\}\ .$$

In other words $\mathcal{G}_{\mathrm{BF}}$ is a collection of sets indexed by Hölder smooth functions of the first $d-1$ coordinates of the feature domain $[0,1]^d$. Therefore the set G^* and the corresponding boundary function g^* are equivalent representations of the Bayes classifier. See Figure 2(b) for an example of such a set. Although the boundary fragment classes might seem artificial and unrealistic they are nevertheless useful in order to determine fundamental performance limits and to gain understanding about more complicated model classes with similar characteristics (e.g., similar characterization of the boundary behavior). There are various examples in the literature where boundary fragments have been used for such purposes, see for example [27], [28].

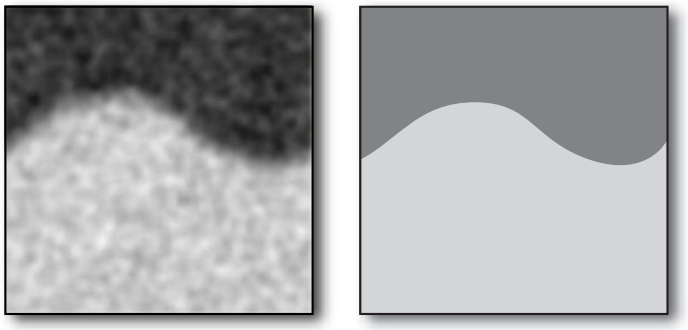
Furthermore, we assume that $p_{\boldsymbol{X}}$, the marginal density of $\boldsymbol{X}$ with respect to the Lebesgue measure, is uniform but, as before, the results in this paper can be easily generalized to the case there $p_{\boldsymbol{X}}$ is bounded above and below, yielding the same rates of error convergence. As in the previous section we require also $\eta(\cdot)$ to have a certain behavior around the decision boundary. Let $\boldsymbol{x} = (\tilde{\boldsymbol{x}}, x_d)$ where $\tilde{\boldsymbol{x}} = (x_1, \ldots, x_{d-1})$. Let $\kappa \geq 1$ and $c > 0$, then for some $\delta > 0$

$$|\eta(\boldsymbol{x}) - 1/2| \geq c|x_d - g^*(\tilde{\boldsymbol{x}})|^{\kappa - 1}\ , \tag{9}$$

for all $\boldsymbol{x}$ such that $|\eta(\boldsymbol{x}) - 1/2| \leq \delta$. The condition above is analogous to the one defined in Section III. We assume as well a reverse-sided condition on $\eta(\cdot)$, namely

$$|\eta(\boldsymbol{x}) - 1/2| \leq C|x_d - g^*(\tilde{\boldsymbol{x}})|^{\kappa - 1}\ , \tag{10}$$

for all $\boldsymbol{x} \in [0,1]^d$, where $C > c$. This condition, together with (9), provides a two-sided characterization of the "noise" around the decision boundary. Let $\mathrm{BF}(\alpha, \kappa, L, c, C)$ be the class of distributions satisfying the noise



(a) (b)

Fig. 2. *(a) Example of the conditional distribution $\eta(\cdot)$ of an element of the class $\mathrm{BF}(\alpha, \kappa, L, c, C)$ when $d = 2$ and $\alpha = 2$. (b) The corresponding Bayes classifier.*

conditions above with parameter κ and whose Bayes classifiers are boundary fragments with smoothness α. An example of such a distribution function, and the corresponding Bayes decision set is presented in Figure 2.

A. Minimax Lower Bounds ($d > 1$)

We begin by presenting lower bounds on the performance of active and passive sampling methods. We start by characterizing active learning for the boundary fragment classes.

Theorem 3. *Let $\rho = (d-1)/\alpha$. Under assumptions (A1) and (A2.2)*

$$\inf_{\widehat{G}_n, S_n} \sup_{P_{\boldsymbol{X}Y} \in \mathrm{BF}(\alpha,\kappa,L,c,C)} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) \geq \mathrm{const}(\alpha, \kappa, L, c, C) n^{-\frac{\kappa}{2\kappa+\rho-2}} \ ,$$

for large enough n, where $\inf_{\widehat{G}_n, S_n}$ denotes the infimum over all possible classifiers and active sampling strategies S_n, and $\mathrm{const}(\alpha, \kappa, L, c, C) > 0$ is a constant factor.

The proof of Theorem 3 is presented in Appendix C. An important remark is that, as before, condition (10) does not play a prominent role in the lower bound, therefore dropping that assumption (equivalently taking $C = \infty$) does not yield slower rates in the theorem statement.

Contrast this result with the one attained for passive sampling: under the passive sampling scenario it is clear that the sample locations $\{\boldsymbol{X}_i\}_{i=1}^n$ must be scattered around the feature space $[0,1]^d$ in a somewhat uniform manner, for example taken uniformly distributed over $[0,1]^d$. A slight modification of the proof of Theorem 3 yields the following proposition.

Proposition 3. *Under assumptions (A1) and (A2.1) we have*

$$\inf_{\widehat{G}_n} \sup_{P_{\boldsymbol{X}Y} \in \mathrm{BF}(\alpha,\kappa,L,c,C)} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) \geq \mathrm{const}(\alpha, \kappa, L, c, C) \cdot n^{-\frac{\kappa}{2\kappa+\rho-1}} \ , \tag{11}$$

where $\rho = (d-1)/\alpha$ and the samples $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. uniformly over $[0,1]^d$.

The rate in this result coincides with the results in [17], although it cannot be derived directly from that work since the model class considered there is larger than the model class considered here. A proof sketch is presented in Appendix D. Furthermore the bound is tight, in the sense that it is possible to devise classification strategies attaining the same asymptotic behavior. We notice that under the passive sampling scenario the excess risk decays at a strictly slower rate than the lower bound for the active sampling scenario, and the rate difference can be dramatic, specially for large smoothness α (equivalently low complexity ρ) and favorable margin (small κ). The active learning lower bound is also tight (in terms of rates, as shown in the next section), which demonstrates that active learning has the potential to improve significantly over passive learning. Finally the result of Theorem 3 is a lower bound, and it therefore applies to the broader classes of distributions introduced in [17], characterized in terms of the metric entropy of the class of Bayes classifiers and a one-sided margin condition, akin to (9).

B. Upper Bounds ($d > 1$)

In this section we present an active learning algorithm for the boundary fragment class and upper bound the corresponding excess risk. The upper bound achieves the rates of Theorem 3 to within a logarithmic factor. This proposed method yields a classifier $\widehat{G}_n$ that has a boundary fragment structure, although the boundary is no longer a smooth function. It is instead a piecewise polynomial function. The methodology proceeds along the lines of [29], [30], extending one-dimensional active sampling results of Section III to this higher dimensional setting. To avoid carrying around cumbersome constants we use the 'big-O' [2] notation for simplicity. Also we use a tilde to denote vectors of dimension $d-1$. We focus the description exclusively on the case $\kappa > 1$, but a similar reasoning gives the results for the case $\kappa = 1$.

We begin by constructing a grid over the first $d-1$ dimensions of the feature domain. Let M be an integer and $\tilde{\boldsymbol{l}} \in \{0, \ldots, M\}^{d-1}$. Define the set of line segments $\mathcal{L}_{\tilde{\boldsymbol{l}}} \stackrel{\Delta}{=} \{(M^{-1}\tilde{\boldsymbol{l}}, x_d) : x_d \in [0,1]\}$. We collect N actively chosen samples along each line, in order to estimate $g(M^{-1}\tilde{\boldsymbol{l}})$. This yields a total of $N(M+1)^{d-1}$ samples (where $n \geq N(M+1)^{d-1}$). We then interpolate the estimates of g at these points to construct a final estimate of the decision boundary. The adequate choices for M and N arise from the performance analysis; for now we point out only that both M and N must be growing with the total number of samples n. Figure 3 illustrates the procedure.

When restricting ourselves to the line segments in $\mathcal{L}_{\tilde{\boldsymbol{l}}}$, the estimation problem boils down to a one-dimensional change-point detection problem and so we can use the results derived in Section III, in particular (8). Since we are using N actively chosen samples per line segment, choosing

$$t = t_N \stackrel{\Delta}{=} c_1 \left(\log N/N\right)^{\frac{1}{2\kappa-2}} \tag{12}$$

guarantees that $\Pr(|\widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) - g^*(M^{-1}\tilde{\boldsymbol{l}})| > t_N) = O\left(N^{-\gamma}\right)$ as $N \to \infty$, where $\gamma > 0$ can be arbitrarily large provided c_1 is sufficiently large.

[2]Let u_n and v_n be two real sequences. We say $u_n = O(v_n)$ as $n \to \infty$ if and only if there exists $C > 0$ and $n_0 > 0$ such that $|u_n| \leq C v_n$ for all $n \geq n_0$.

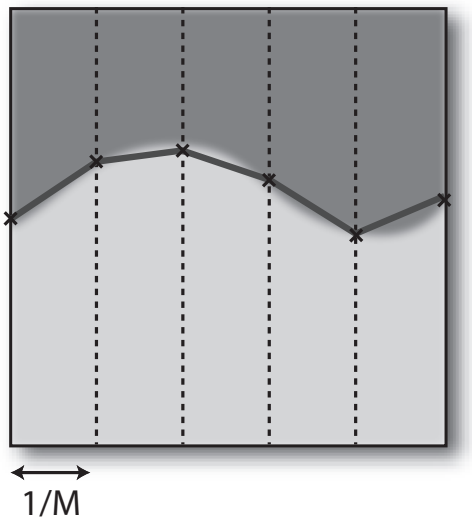


Fig. 3. *Illustration of the active classification procedure for boundary fragments when $d = 2$. In this case $\alpha = 2$ therefore we estimate the true Bayes decision boundary with the aid of piecewise linear polynomials. The crosses represent the estimates of g^* obtained by each one of the $M + 1$ line searches. The dark solid line segments represent the $M/\lceil\alpha\rceil$ interpolation polynomials.*

Let $\{\widehat{g}(M^{-1}\tilde{\boldsymbol{l}})\}$ be the estimates obtained using this method at each of the points indexed by $\tilde{\boldsymbol{l}}$. We use these estimates to construct a piecewise polynomial fit to approximate g^*. In what follows assume $\alpha > 1$. The case $\alpha = 1$ can be handled in a very similar way (where one would approximate g^* with a stair function). Begin by dividing $[0,1]^{d-1}$ (that is, the domain of g^*) into cells. Without loss of generally assume that M is such that $M/\lfloor\alpha\rfloor$ is an integer (this can be enforced with the proper choice of M). Let $\tilde{\boldsymbol{q}} \in \{0,\ldots,M/\lfloor\alpha\rfloor - 1\}^{d-1}$ index the cells

$$I_{\tilde{\boldsymbol{q}}} \triangleq \left[\tilde{\boldsymbol{q}}_1\lfloor\alpha\rfloor M^{-1}, (\tilde{\boldsymbol{q}}_1+1)\lfloor\alpha\rfloor M^{-1}\right] \times \cdots \times \left[\tilde{\boldsymbol{q}}_{d-1}\lfloor\alpha\rfloor M^{-1}, (\tilde{\boldsymbol{q}}_{d-1}+1)\lfloor\alpha\rfloor M^{-1}\right] \ .$$

Note that these cells partition the domain $[0,1]^{d-1}$ entirely. In each one of the cells we perform a polynomial interpolation using the estimates of g^* at points within the cell. For the cell indexed by $\tilde{\boldsymbol{q}}$ we consider a tensor product polynomial fit $\widehat{L}_{\tilde{\boldsymbol{q}}}$, that can be written as

$$\widehat{L}_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) = \sum_{\tilde{\boldsymbol{l}}:M^{-1}\tilde{\boldsymbol{l}}\in I_{\tilde{\boldsymbol{q}}}} \widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \ ,$$

where $\tilde{\boldsymbol{x}} \in [0,1]^{d-1}$. The functions $Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}$ are the tensor-product Lagrange polynomials (see for example [31]),

$$Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \triangleq \prod_{i=1}^{d-1} \prod_{j=0, j\neq\tilde{\boldsymbol{l}}_i - \lfloor\alpha\rfloor\tilde{\boldsymbol{q}}_i}^{\lfloor\alpha\rfloor} \frac{\tilde{\boldsymbol{x}}_i - M^{-1}(\lfloor\alpha\rfloor\tilde{\boldsymbol{q}}_i + j)}{M^{-1}\tilde{\boldsymbol{l}}_i - M^{-1}(\lfloor\alpha\rfloor\tilde{\boldsymbol{q}}_i + j)} \ .$$

We remark that this is a polynomial interpolation and so we have that $\widehat{L}_{\tilde{\boldsymbol{q}}}(M^{-1}\tilde{\boldsymbol{l}}) = \widehat{g}(M^{-1}\tilde{\boldsymbol{l}})$, for all $\tilde{\boldsymbol{l}}$ such that $M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}$. Finally, the estimate $\widehat{g}$ of g^* is given by

$$\widehat{g}(\tilde{\boldsymbol{x}}) = \sum_{\tilde{\boldsymbol{q}}\in\{0,\ldots,M/\lfloor\alpha\rfloor-1\}^{d-1}} \widehat{L}_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}$$

which defines a classification rule $\widehat{G}_n$.

Theorem 4. *Let* $M = \lfloor \alpha \rfloor \left\lfloor n^{\frac{1}{\alpha(2\kappa-2)+d-1}} \right\rfloor$ *and* $N = \lfloor n/(M+1)^{d-1} \rfloor$*, and consider the active learning algorithm described above. Let* $\rho = (d-1)/\alpha$*, then*

$$\sup_{P_{\boldsymbol{X}Y} \in \mathrm{BF}(\alpha,\kappa,L,c,C)} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) = O\left((\log n/n)^{\frac{\kappa}{2\kappa+\rho-2}} \right) .$$

The proof of Theorem 4 is given in Appendix E. One sees that this estimator achieves the rate of Theorem 3 to within a logarithmic factor. It is not clear if the logarithmic factor is an artifact of our construction, or if it is unavoidable. One knows [29] that if $\kappa, \alpha = 1$ the logarithmic factor can be eliminated by using a slightly more sophisticated interpolation scheme.

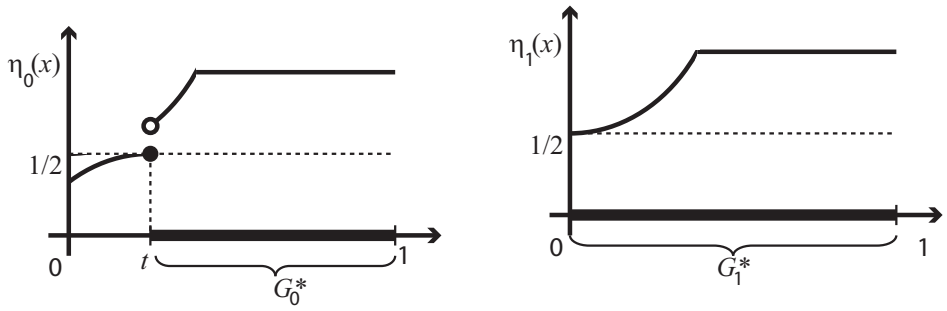
V. Final Remarks

This paper presented upper and lower bounds for active learning algorithms under assumptions on the decision boundary regularity and noise conditions. Since the upper and lower bounds agree up to a logarithmic factor, we may conclude that lower bound is near minimax optimal. That is, for the distributional classes under consideration, no active or passive learning procedure can perform significantly better in terms of excess risk decay rates. The upper bounds were derived constructively, based on active learning algorithms originally developed for one-dimensional change-point detection. Note that the noise characterization in (2) and (9) is key to guarantee that active learning performs better than passive learning in a minimax sense, by enforcing that there is a strong enough cue active learning methods can use to focus on the decision boundary.

In principle, the methodology employed in the upper bound calculation could be applied in practice in the case of boundary fragments and with knowledge of the key regularity parameters κ and ρ. Unfortunately this is not a scenario one expects to have in practice, and thus a key open problem is the design of active learning algorithms that are adaptive to unknown regularity parameters and capable of handling arbitrary boundaries (not only fragments). Nonetheless, the results of this paper do indicate fundamental limitations of active learning, and thus can provide guidelines for performance expectations in practice. Moreover, the bounds clarify the situations in which active learning can lead to significant gains over passive learning, and it may be possible to assess the conditions that might hold in a given application in order to gauge the merit of pursuing an active learning approach. One practical algorithm that exhibits near optimal rates for more general boundaries in the bounded noise case is the multiscale technique proposed in [10], and it may be possible to devise similar methods for more general noise conditions.

Acknowledgements

The authors would like to thank Aarti Singh and two anonymous reviewers for carefully reading the manuscript and providing valuable feedback, leading to a much clearer presentation.



(a) (b)

Fig. 4. *The two conditional distributions used for the proof of Theorem 1.*

APPENDIX A
PROOF OF THEOREM 1

The proof strategy follows the basic idea behind standard minimax analysis methods, and consists in reducing the problem of classification in the class $\mathcal{P}(\kappa, c, C)$ to a hypothesis testing problem. In this case it suffices to consider two hypothesis and use the following result from [25] (page 76, theorem 2.2).

Theorem 5 (Tsybakov 2004)**.** *Let $\mathcal{F}$ be a class of models. Associated with each model $f \in \mathcal{F}$ we have a probability measure P_f defined on a common probability space. Let $d(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a semi-distance. Let $f_0, f_1 \in \mathcal{F}$ be such that $d(f_0, f_1) \geq 2a$, with $a > 0$. Assume also that $\mathrm{KL}(P_{f_1} \| P_{f_0}) \leq \gamma$, where KL denotes the Kullback-Leibler divergence*[3] *The following bound holds.*

$$\begin{aligned} \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} P_f \left(d(\widehat{f}, f) \geq a \right) &\geq \inf_{\widehat{f}} \max_{j \in \{0,1\}} P_{f_j} \left(d(\widehat{f}, f_j) \geq a \right) \\ &\geq \max \left(\frac{1}{4} \exp(-\gamma), \frac{1 - \sqrt{\gamma/2}}{2} \right) , \end{aligned}$$

where the infimum is taken with respect to the collection of all possible estimators of f (based on a sample from P_f).

To prove the statement of Theorem 1 we take $\mathcal{F} = \mathcal{P}(\kappa, c, C)$ and are interested in controlling the excess risk

$$R_P(\widehat{G}_n) - R_P(G_P^*) = \int_{\widehat{G}_n \Delta G_P^*} |2\eta_P(x) - 1| \mathrm{d}x \ ,$$

[3]Let P and Q be two probability measures defined on a common probability space $(\Omega, \mathcal{B})$. Then P is dominated by Q, that is $P \ll Q$, if and only if for all $B \in \mathcal{B}$, $Q(B) = 0 \Rightarrow P(B) = 0$. 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.

where the subscript P indicates that the excess risk is being measured with respect to the distribution $P \in \mathcal{P}(\kappa, c, C)$. Since the excess risk is not a semi-distance we cannot apply Theorem 5 directly, but we can relate excess risk and the symmetric distance measure, and then use the theorem. The two distributions/hypotheses we consider are completely characterized by the conditional probability η (since the marginal distribution of X is uniform). Let

$$\eta_0(x) = \begin{cases} \min\left(\frac{1}{2} + c\,\mathrm{sign}(x-t)|x-t|^{\kappa-1}, 1\right) & , x \le A \\ \min\left(\frac{1}{2} + c\,x^{\kappa-1}, 1\right) & , x > A \end{cases} , \tag{13}$$

$$\eta_1(x) = \min\left(\frac{1}{2} + c\,x^{\kappa-1}, 1\right) , \tag{14}$$

where the appropriate value t is going to be chosen later and $A = t\left(1 + \frac{1}{(C/c)^{1/(\kappa-1)}-1}\right)$. Assume $t \le 1/(2c)^{1/(\kappa-1)}$ so that $\eta_0(\cdot)$ is a proper conditional probability function. These functions are depicted in Figure 4 when $C = \infty$. Note that $G_0^* = [t, 1]$ and $G_1^* = [0, 1]$ (provided t is small enough). In what follows we use the subscript 0 or 1 whenever we want to denote explicitly the dependence on the underlying model (respectively characterized by η_0 or η_1). Begin by observing that the two constructed distribution belong to the class $\mathcal{F} = \mathcal{P}(\kappa, c, C)$ of interest, that is, these two distributions satisfy the margin conditions (2) and (3). Another key observation is the relationship between the symmetric difference and the excess risk. For any set $G \subseteq [0, 1]$ and $j \in \{0, 1\}$ we have

$$R_j(G) - R_j(G_j^*) \ge \min\left\{\frac{4c}{\kappa 2^\kappa} d_\Delta(G, G_j^*)^\kappa, d_\Delta(G, G_j^*)\right\} , \tag{15}$$

where $d_\Delta(G, G_j^*) \triangleq \int_{G \Delta G_j^*} \mathrm{d}x$ is the symmetric difference semi-distance. Note that when the symmetric difference is small the first term in between the curly brackets dominates. To show (15) consider the case $j = 0$ (the case $j = 1$ is analogous). Let G be such that $d_\Delta(G, G_0^*) = \tau$. Then

$$\begin{aligned} R_0(G) - R_0(G_0^*) = \int_{G \Delta G_0^*} |2\eta_0(x) - 1| \mathrm{d}x &\ge \int_{G \Delta G_0^*} \min\{2c|t-x|^{\kappa-1}, 1\} \mathrm{d}x \\ &\ge \min\left\{\int_{t-\tau/2}^{t+\tau/2} 2c|t-x|^{\kappa-1} \mathrm{d}x, \tau\right\} \\ &= \min\left\{2\int_{t}^{t+\tau/2} 2c(x-t)^{\kappa-1} \mathrm{d}x, \tau\right\} \\ &= \min\left\{\frac{4c}{\kappa 2^\kappa}\tau^\kappa, \tau\right\} . \end{aligned}$$

We now proceed by applying Theorem 5 to the semi-distance d_Δ and posteriorly use (15) to control the excess risk. Note that $d_\Delta(G_0^*, G_1^*) = t$. Let $P_{0,n} \triangleq P^{(0)}_{X_1,\ldots,X_n,Y_1,\ldots,Y_n}$ be the probability measure of the random variables $\{X_i, Y_i\}_{i=1}^n$ under hypothesis 0 and define analogously $P_{1,n} \triangleq P^{(1)}_{X_1,\ldots,X_n,Y_1,\ldots,Y_n}$. Define $\boldsymbol{Z}_j^X \triangleq (X_1, \ldots, X_j)$ and

$\boldsymbol{Z}_j^Y \triangleq (Y_1, \ldots, Y_j)$. Then

$$\begin{aligned}
\mathrm{KL}(P_{1,n} \| P_{0,n}) &= \mathbb{E}_1 \left[\log \frac{P^{(1)}_{\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y}(\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y)}{P^{(0)}_{\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y}(\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y)} \right] \\
&= \mathbb{E}_1 \left[\log \frac{\prod_{j=1}^n P^{(1)}_{Y_j|X_j}(Y_j|X_j)\ P_{X_j|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y}(X_j|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y)}{\prod_{j=1}^n P^{(0)}_{Y_j|X_j}(Y_j|X_j)\ P_{X_j|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y}(X_j|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y)} \right] \qquad (16) \\
&= \mathbb{E}_1 \left[\log \frac{\prod_{j=1}^n P^{(1)}_{Y_j|X_j}(Y_j|X_j)}{\prod_{j=1}^n P^{(0)}_{Y_j|X_j}(Y_j|X_j)} \right] \\
&= \sum_{j=1}^n \mathbb{E}_1 \left[\mathbb{E}_1 \left[\log \frac{P^{(1)}_{Y_j|X_j}(Y_j|X_j)}{P^{(0)}_{Y_j|X_j}(Y_j|X_j)} \middle| X_1, \ldots, X_n \right] \right] \\
&\leq n \max_{x \in [0,1]} \mathbb{E}_1 \left[\log \frac{P^{(1)}_{Y_1|X_1}(Y_1|X_1)}{P^{(0)}_{Y_1|X_1}(Y_1|X_1)} \middle| X_1 = x \right] ,
\end{aligned}$$

where in the above $\mathbb{E}_1$ denotes the expectation taken with respect to measure $P_{1,n}$. Step (16) follows since the distribution of X_j conditional on $\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y$ depends only on the sampling strategy S_n, and does not change with the underlying distribution, therefore those terms in the numerator and denominator cancel out. The last step follows from the observation that, conditional on the feature vectors, the labels Y_j are independent and identically distributed. The expectation in the last line is the Kullback-Leibler divergence between two Bernoulli random-variables. The following straightforward result provides a bound on that divergence.

Lemma 1. *Let P and Q be Bernoulli random variables with parameters respectively $1/2 - p$ and $1/2 - q$. Let $|p|, |q| \leq 1/4$, then $\mathrm{KL}(P\|Q) \leq 8(p-q)^2$.*

We conclude that

$$\begin{aligned}
\mathrm{KL}(P_{1,n} \| P_{0,n}) &\leq \max_{x \in [0,1]} 8n(\eta_1(x) - \eta_0(x))^2 \leq 8nc^2 \max_{x \in [0,1]} (|x-t|^{\kappa-1} + |x|^{\kappa-1})^2 \\
&\leq 8nc^2 (2A^{\kappa-1})^2 = 32c^2 \left(1 + \frac{1}{(C/c)^{1/(\kappa-1)} - 1} \right)^{2\kappa-2} nt^{2\kappa-2} \\
&= c_0 \cdot nt^{2\kappa-2} ,
\end{aligned}$$

provided A (consequently t) is small enough so that $|\eta_i(A) - 1/2| \leq 1/4,\ i \in \{0,1\}$. In the above expression we have $c_0 \triangleq 32c^2 \left(1 + \frac{1}{(C/c)^{1/(\kappa-1)} - 1} \right)^{2\kappa-2}$, a constant factor.

Taking $t = n^{-\frac{1}{2\kappa-2}}$ and using Theorem 5 we conclude that for n large enough (implying t small).

$$\inf_{\widehat{G}_n} \max_{j \in \{0,1\}} P_j \left(d_\Delta(\widehat{G}_n, G_j^*) \geq t/2 \right) \geq \frac{1}{4} \exp(-c_0) > 0 ,$$

We can now use (15) to conclude that $P_j \left(R_j(G) - R_j(G_j^*) \geq \min \left\{ \frac{4c}{\kappa 2^\kappa} (t/2)^\kappa, t/2 \right\} \right) \geq P_j(d_\Delta(G, G_j^*) \geq t/2)$.

Therefore, taking n so that t is small we have

$$\begin{aligned}\inf_{\widehat{G}_n}\sup_{P_{XY}\in\mathcal{P}(\kappa,c,C)} & P\left(R(\widehat{G}_n)-R(G^*)\geq\frac{4c}{\kappa 4^k}\cdot n^{-\frac{\kappa}{2\kappa-2}}\right)\\ \geq & \inf_{\widehat{G}_n}\max_{j\in\{0,1\}} P_j\left(R(\widehat{G}_n)-R(G^*)\geq\frac{4c}{\kappa 4^k}\cdot n^{-\frac{\kappa}{2\kappa-2}}\right)\\ \geq & \inf_{\widehat{G}_n}\max_{j\in\{0,1\}} P_j(d_\Delta(G,G_j^*)\geq t/2)\geq\frac{1}{4}\exp(-c_0)>0\ ,\end{aligned}$$

The statement of the theorem now follows from the application of Markov's inequality to the above expression,

$$\mathbb{E}\left[R(\widehat{G}_n)-R(G^*)\right] \geq \frac{4c}{\kappa 4^k}n^{-\frac{\kappa}{2\kappa-2}}P\left(R(\widehat{G}_n)-R(G^*)\geq\frac{4c}{\kappa 4^k}n^{-\frac{\kappa}{2\kappa-2}}\right)\ .$$

■

APPENDIX B
SKETCH PROOF OF PROPOSITION 1

The proof is nearly identical to the proof of Theorem 1, differing only on the bounding of the Kullback-Leibler divergence between the two hypotheses. When working under the passive sampling assumption (A2.1) the best sampling scheme needs to place the samples somewhat uniformly over $[0,1]$. If this is not the case one can construct two hypotheses that are "hard to distinguish" under this sampling scheme (that is, the KL divergence $\mathrm{KL}(P_{1,n}\|P_{0,n})$ is small), yielding an incorrect worst case bound. Without loss of generality the best passive sampling strategy for the problem at hand takes X_i i.i.d. uniformly over $[0,1]$. In such a scenario we can use the construction in Appendix A and the Kullback divergence $\mathrm{KL}(P_{1,n}\|P_{0,n})$ is approximately proportional to $nt^{2\kappa-2}\cdot t = nt^{2\kappa-1}$, since roughly only a fraction $A\sim t$ of the samples are informative (any sample taken in $(A,1]$ is non-informative). Taking $t\sim n^{-1/(2\kappa-1)}$ and proceeding as before yields the passive sampling minimax bound (4). ■

APPENDIX C
PROOF OF THEOREM 3

As the proof of Theorem 1, the following proof uses standard techniques for the most part, but to get the bounds desired we need now more than only two hypotheses. The main tool is the following theorem, from [25] (page 85, theorem 2.5).

Theorem 6 (Tsybakov, 2004)**.** *Let $\mathcal{F}$ be a class of models. Associated with each model $f\in\mathcal{F}$ we have a probability measure P_f defined on a common probability space. Let $M\geq 2$ be an integer and let $d(\cdot,\cdot):\mathcal{F}\times\mathcal{F}\to\mathbb{R}$ be a collection of semi-distances. Suppose we have $\{f_0,\ldots,f_M\}\in\mathcal{F}$ such that*

i) $d(f_j,f_k)\geq 2\,a>0,\quad \forall_{0\leq j,k\leq M}$,

ii) $P_{f_0}\ll P_{f_j},\quad \forall_{j=1,\ldots,M}$, *(for notation clarification see footnote 3 on page 19)*

iii) $\frac{1}{M}\sum_{j=1}^M \mathrm{KL}(P_{f_j}\|P_{f_0})\leq\gamma\log M$, *where* $0<\gamma<1/8$.

The following bound holds.

$$\begin{aligned}\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} P_f\left(d(\widehat{f},f)\geq a\right) &\geq \inf_{\widehat{f}} \max_{j\in\{0,\ldots,M\}} P_{f_j}\left(d(\widehat{f},f_j)\geq a\right)\\ &\geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\gamma-2\sqrt{\frac{\gamma}{\log M}}\right)>0\ ,\end{aligned}$$

where the infimum is taken with respect to the collection of all possible estimators of f (based on a sample from P_f).

As in the proof of Theorem 1 we are going to construct a bound on the performance of an estimator measured according to d_Δ. By later relating this semi-distance with the excess risk we obtain the desired lower bound. To apply the theorem we need to construct finite subset of distributions in $\mathrm{BF}(\alpha,\kappa,L,c,C)$. The elements of this set are distributions $P_{\boldsymbol{X}Y}$ and therefore uniquely characterized by the conditional probability $\eta(\boldsymbol{x})=\Pr(Y=1|\boldsymbol{X}=\boldsymbol{x})$ (since we are assuming that $P_{\boldsymbol{X}}$ is uniform over $[0,1]^d$). Let $\boldsymbol{x}=(\tilde{\boldsymbol{x}},x_d)$ with $\tilde{\boldsymbol{x}}\in[0,1]^{d-1}$. As a notational convention we use a tilde to denote a vector of dimension $d-1$. Define

$$m=\left\lceil c_0 n^{\frac{1}{\alpha(2\kappa-2)+d-1}}\right\rceil\ ,\quad \tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}=\frac{\tilde{\boldsymbol{l}}-1/2}{m}\ ,$$

where $\tilde{\boldsymbol{l}}\in\{1,\ldots,m\}^{d-1}$ and $c_0>0$ is to be determined later. Define also $\varphi_{\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}})=Lm^{-\alpha}h(m(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}))$, with $h\in\Sigma(1,\alpha)$, $\mathrm{supp}(h)=(-1/2,1/2)^{d-1}$ and $h\geq 0$. It is easily shown that such a function exists, for example

$$h(\tilde{\boldsymbol{x}})=a\prod_{i=1}^{d-1}\exp\left(-\frac{1}{1-4x_i^2}\right)\mathbf{1}\{|x_i|<1/2\}\ ,$$

with $a>0$ sufficiently small. The functions $\varphi_{\tilde{\boldsymbol{l}}}$ are little "bumps" centered at the points $\tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}$. The collection $\{\tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}\}$ forms a regular grid over $[0,1]^{d-1}$.

Let $\Omega=\{\boldsymbol{\omega}=(\omega_1,\ldots,\omega_{m^{d-1}}),\ \omega_i\in\{0,1\}\}=\{0,1\}^{m^{d-1}}$, and define

$$g_{\boldsymbol{\omega}}(\cdot)=\sum_{\tilde{\boldsymbol{l}}\in\{1,\ldots,m\}^{d-1}}\omega_{\tilde{\boldsymbol{l}}}\ \varphi_{\tilde{\boldsymbol{l}}}(\cdot),\ \boldsymbol{\omega}\in\Omega\ .$$

The functions $g_{\boldsymbol{\omega}}$ are boundary functions. The binary vector $\boldsymbol{\omega}$ is an indicator vector: if $\omega_{\tilde{\boldsymbol{l}}}=1$ then "bump" $\tilde{\boldsymbol{l}}$ is present, otherwise that "bump" is absent. Note that $\varphi_{\tilde{\boldsymbol{l}}}\in\Sigma(L,\alpha)$ and these functions have disjoint support, therefore $g_{\boldsymbol{\omega}}\in\Sigma(L,\alpha)$ for all $\boldsymbol{\omega}\in\Omega$. Let $\boldsymbol{\omega}\in\Omega$ and construct the conditional distribution

$$\eta_{\boldsymbol{\omega}}(\boldsymbol{x})=\begin{cases}\min\left(\frac{1}{2}+c\cdot\mathrm{sign}(x_d-g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}}))|x_d-g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}})|^{\kappa-1},1\right), & \text{if } x_d\leq A\\ \min\left(\frac{1}{2}+c\cdot x_d^{\kappa-1},1\right), & \text{if } x_d>A\end{cases}\ ,$$

$$A=\max_{\tilde{\boldsymbol{x}}}\varphi(\tilde{\boldsymbol{x}})\left(1+\frac{1}{(C/c)^{1/(\kappa-1)}-1}\right)=Lm^{-\alpha}h_{\max}\left(1+\frac{1}{(C/c)^{1/(\kappa-1)}-1}\right)\ ,$$

with $h_{\max}=\max_{\tilde{\boldsymbol{x}}\in\mathbb{R}^{d-1}}h(\tilde{\boldsymbol{x}})$. The choice of A is done carefully, in order to ensure that the functions $\eta_{\boldsymbol{\omega}}$ are similar, but at the same time satisfy the margin conditions. It is easily checked that conditions (9) and (10) are satisfied

for the distributions above. By construction the Bayes decision boundary for each of these distributions is given by $x_d = g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}})$ and so these distributions belong to the class $\mathrm{BF}(\alpha, \kappa, L, c, C)$. Note also that these distributions are all identical if $x_d > A$. As n increases m also increases and therefore A decreases, so the conditional distributions described above are becoming more and more similar. This is key to bound the Kullback-Leibler divergence between these distributions.

The above collection of distributions, indexed by $\boldsymbol{\omega} \in \Omega$, is still too large for the application of Theorem 6. Recall the following lemma.

Lemma 2 (Varshamov-Gilbert bound, 1962)**.** *Let $m^{d-1} \geq 8$. There exists a subset $\{\boldsymbol{\omega}^{(0)}, \boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)}\}$ of Ω such that $M \geq 2^{m^{d-1}/8}$, $\boldsymbol{\omega}^{(0)} = (0, \ldots, 0)$ and*

$$\rho(\boldsymbol{\omega}^{(j)}, \boldsymbol{\omega}^{(k)}) \geq m^{d-1}/8, \quad \forall\ 0 \leq j < k \leq M\ ,$$

where ρ denotes the Hamming distance.

For a proof of the Lemma 2 see [25](page 89, lemma 2.7). To apply Theorem 6 we use the M distributions $(\eta_{\boldsymbol{\omega}^{(0)}}, \ldots, \eta_{\boldsymbol{\omega}^{(M)}})$ given by the lemma. For each distribution $\eta_{\boldsymbol{\omega}^{(i)}}$ we have the corresponding Bayes classifier G_i^*. As before recall that $d_\Delta(G, G') = \int_{G \Delta G'} \mathrm{d}\boldsymbol{x}$.

Let $i \neq j$. By construction we observe that

$$\begin{aligned}
d_\Delta(G_j^*, G_i^*) &= \int_{[0,1]^{d-1}} \int_0^{|g_i^*(\tilde{\boldsymbol{x}}) - g_j^*(\tilde{\boldsymbol{x}})|} 1 \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\
&= \sum_{\tilde{\boldsymbol{l}} \in \{1,\ldots,m\}^{d-1}} |\omega_{\tilde{\boldsymbol{l}}}^{(i)} - \omega_{\tilde{\boldsymbol{l}}}^{(j)}| \int_{[0,1]^{d-1}} \int_0^{\varphi_{\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}})} 1\ \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \qquad (17) \\
&= \sum_{\tilde{\boldsymbol{l}} \in \{1,\ldots,m\}^{d-1}} |\omega_{\tilde{\boldsymbol{l}}}^{(i)} - \omega_{\tilde{\boldsymbol{l}}}^{(j)}| \int_{[0,1]^{d-1}} L m^{-\alpha} h(m(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}})) \mathrm{d}\tilde{\boldsymbol{x}} \\
&= \sum_{\tilde{\boldsymbol{l}} \in \{1,\ldots,m\}^{d-1}} |\omega_{\tilde{\boldsymbol{l}}}^{(i)} - \omega_{\tilde{\boldsymbol{l}}}^{(j)}| \int_{[-1/2,1/2]^{d-1}} L m^{-\alpha-(d-1)} h(\tilde{\boldsymbol{z}}) \mathrm{d}\tilde{\boldsymbol{z}} \qquad (18) \\
&= \sum_{\tilde{\boldsymbol{l}} \in \{1,\ldots,m\}^{d-1}} |\omega_{\tilde{\boldsymbol{l}}}^{(i)} - \omega_{\tilde{\boldsymbol{l}}}^{(j)}| L m^{-\alpha-(d-1)} \|h\|_1 \\
&\geq \rho(\boldsymbol{\omega}^{(i)}, \boldsymbol{\omega}^{(j)}) L m^{-\alpha-(d-1)} \|h\|_1 \\
&\geq \frac{L\|h\|_1}{8} m^{-\alpha}\ ,
\end{aligned}$$

where $\|h\|_1$ denotes the L_1 norm of h. In the above step (17) follows from the fact that $\varphi_{\tilde{\boldsymbol{l}}}(\cdot)$ and $\varphi_{\tilde{\boldsymbol{k}}}(\cdot)$ have disjoint support provided $\tilde{\boldsymbol{l}} \neq \tilde{\boldsymbol{k}}$, and step (18) is due to a simple linear change of variable (which introduces the factor $m^{-(d-1)}$).

The next step of the proof is to lower-bound $R_i(G) - R_i(G_i^*)$ using $d_\Delta(G, G_i^*)$, where $G \subseteq [0,1]^d$. Suppose $d_\Delta(G, G_i^*) = \tau$. The smallest excess risk $R_i(G) - R_i(G_i^*)$ is attained when the points in set G coincide with the

points $\tilde{\boldsymbol{x}}$ such that $\eta_{\boldsymbol{\omega}^{(i)}}$ is closest to $1/2$. Taking this into account we observe that

$$\begin{aligned} R_i(G) - R_i(G_i^*) &\geq \int_{[0,1]^{d-1}} \int_{g_i^*(\tilde{\boldsymbol{x}})-\tau/2}^{g_i^*(\tilde{\boldsymbol{x}})+\tau/2} \min\left\{2c|x_d - g_i^*(\tilde{\boldsymbol{x}})|^{\kappa-1}, 1\right\} \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\ &= \int_{[0,1]^{d-1}} 2\int_{g_i^*(\tilde{\boldsymbol{x}})}^{g_i^*(\tilde{\boldsymbol{x}})+\tau/2} \min\left\{2c(x_d - g_i^*(\tilde{\boldsymbol{x}}))^{\kappa-1}, 1\right\} \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\ &= 2\int_{[0,1]^{d-1}} \int_0^{\tau/2} \min\left\{2cz^{\kappa-1}, 1\right\} \mathrm{d}z \mathrm{d}\tilde{\boldsymbol{x}} = \min\left\{\frac{4c}{\kappa 2^\kappa}\tau^\kappa, \tau\right\} . \end{aligned}$$

Therefore we conclude that for all $G \subseteq [0,1]^d$ we have

$$R_i(G) - R_i(G_i^*) \geq \min\left\{\frac{4c}{\kappa 2^\kappa} d_\Delta^\kappa(G, G_i^*), d_\Delta^\kappa(G, G_i^*)\right\} . \tag{19}$$

We are ready for the final step of the proof. Now let P_i be the distribution of $(\boldsymbol{X}_1, Y_1, \ldots, \boldsymbol{X}_n, Y_n)$ assuming the underlying conditional distribution is $\eta_{\boldsymbol{\omega}^{(i)}}$. We proceed like in the proof of Theorem 1.

$$\begin{aligned} \mathrm{KL}(P_i \| P_0) &\leq n \max_{\boldsymbol{x}\in[0,1]^d} \mathbb{E}_i\left[\log \frac{P^{(i)}_{Y_1|\boldsymbol{X}_1}(Y_1|\boldsymbol{X}_1)}{P^{(0)}_{Y_1|\boldsymbol{X}_1}(Y_1|\boldsymbol{X}_1)}\middle| \boldsymbol{X}_1 = \boldsymbol{x}\right] \\ &\leq 8n(2cA^{\kappa-1})^2 = c_1 \cdot nm^{-\alpha(2\kappa-2)} , \end{aligned}$$

where $c_1 > 0$ and the last inequality holds provided n is large enough so that A is small enough and Lemma 1 can be applied. Finally

$$\frac{1}{M}\sum_{i=1}^M \mathrm{KL}(P_i\|P_0) \leq c_1 \cdot nm^{-\alpha(2\kappa-2)} \leq c_1\ c_0^{-(\alpha(2\kappa-2)+d-1)} m^{d-1} .$$

From Lemma 2 we also have $\frac{\gamma}{8}m^{d-1}\log 2 \leq \gamma \log M$ therefore choosing c_0 large enough in the definition of m guarantees the conditions of Theorem 6 and so

$$\begin{aligned} &\inf_{\widehat{G}_n, S_n} \max_{j\in\{0,\ldots,M\}} P_j\left(d_\Delta(\widehat{G}_n, G_j^*) \geq \frac{L\|h\|_1}{16} m^{-\alpha}\right) \geq \\ &\qquad \inf_{\widehat{G}_n, S_n} \max_{j\in\{0,\ldots,M\}} P_j\left(d_\Delta(\widehat{G}_n, G_j^*) \geq \frac{L\|h\|_1 c_0^{-\alpha}}{16} n^{-\frac{1}{2\kappa-2+(d-1)/\alpha}}\right) \geq c_2 , \end{aligned}$$

for n large enough, where $c_2 > 0$ comes from Theorem 6. Now using (19) similarly to the proof of Theorem 1 we obtain

$$\inf_{\widehat{G}_n, S_n} \sup_{P\in \mathrm{BF}(\alpha,\kappa,L,c,C)} P\left(R(\widehat{G}_n) - R(G^*) \geq \frac{c}{4\kappa 2^\kappa} L\|h\|_1 c_0^{-\alpha} \cdot n^{-\frac{\kappa}{2\kappa-2+(d-1)/\alpha}}\right) \geq c_2 ,$$

provided n is large enough. An application of Markov's inequality yields the original statement of the theorem, concluding the proof.

APPENDIX D
PROOF OF PROPOSITION 3

As in the proof of Proposition 1 one just needs to modify slightly the proof of Theorem 3. Note that if the passive sampling scenario is considered the sample locations $\{\boldsymbol{X}_i\}_{i=1}^n$ have to be selected before any observations are made, therefore they must be somewhat uniformly distributed over the feature domain, in this case $[0,1]^d$. Using the same reasoning of Appendix B we have that $\mathrm{KL}(P_i\|P_0) \leq 8n(cA^{\kappa-1})^2 L h_{\max} \cdot A \sim nm^{-2\alpha(\kappa-1)}m^{-\alpha} \sim nm^{-\alpha(2\kappa-1)}$, since only a fraction $A \sim m^{-\alpha}$ of the samples are informative. Therefore choosing $m \sim n^{\frac{1}{\alpha(2\kappa-1)+d-1}}$ and proceeding in analogous fashion as before yields bound (11).

■

APPENDIX E
PROOF OF THEOREM 4

The proof methodology aims at controlling the excess risk for an event that happens with high probability. To avoid carrying around cumbersome constants we use the 'big-O' notation. We show the proof only for the case $\kappa > 1$, since the proof when $\kappa = 1$ is almost analogous.

Define the event $\Omega_n = \left\{\forall \tilde{\boldsymbol{l}} \in \{0,\ldots,M\}^{d-1} \quad |\widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) - g^*(M^{-1}\tilde{\boldsymbol{l}})| \leq t_N\right\}$. In words, Ω_n is the event that the M^{d-1} point estimates of g do not deviate very much from the true values. Using a union bound, taking into account (8) and the choice t_N in (12) one sees that $1 - \Pr(\Omega_n) = O(N^{-\gamma}M^{d-1})$, where γ can be chosen arbitrarily large. With the choice of M in the theorem and choosing c_1 wisely in the definition of t_N (12) we have $1 - \Pr(\Omega_n) = O\left(n^{-\frac{\alpha\kappa}{\alpha(2\kappa-2)+d-1}}\right)$.

The excess risk of our classifier is given by

$$\begin{aligned}
R(\widehat{G}_n) - R(G^*) &= \int_{\widehat{G}_n \Delta G^*} |2\eta(\boldsymbol{x}) - 1| \mathrm{d}\boldsymbol{x} \\
&= \int_{[0,1]^{d-1}} \int_{\min(\widehat{g}(\tilde{\boldsymbol{x}}),g^*(\tilde{\boldsymbol{x}}))}^{\max(\widehat{g}(\tilde{\boldsymbol{x}}),g^*(\tilde{\boldsymbol{x}}))} |2\eta((\tilde{\boldsymbol{x}}, x_d)) - 1| \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\
&\leq \int_{[0,1]^{d-1}} \int_{\min(\widehat{g}(\tilde{\boldsymbol{x}}),g^*(\tilde{\boldsymbol{x}}))}^{\max(\widehat{g}(\tilde{\boldsymbol{x}}),g^*(\tilde{\boldsymbol{x}}))} 2C|x_d - g(\tilde{\boldsymbol{x}})|^{\kappa-1} \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\
&= 2C \int_{[0,1]^{d-1}} \int_0^{|\widehat{g}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})|} z^{\kappa-1} \mathrm{d}z \mathrm{d}\tilde{\boldsymbol{x}} \\
&= \frac{2C}{\kappa} \int_{[0,1]^{d-1}} |\widehat{g}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})|^{\kappa} \mathrm{d}\tilde{\boldsymbol{x}} = O\left(\|\widehat{g} - g^*\|_\kappa^\kappa\right) \ ,
\end{aligned}$$

where the inequality follows from condition (10), and $\|\cdot\|_\kappa$ denotes the L_κ norm of a function.

Let $L_{\tilde{\boldsymbol{q}}}$, $\tilde{\boldsymbol{q}} \in \{0,\ldots,M/\lfloor\alpha\rfloor - 1\}^{d-1}$ be a clairvoyant version of $\widehat{L}_{\tilde{\boldsymbol{q}}}$, that is,

$$L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) = \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} g^*(M^{-1}\tilde{\boldsymbol{l}}) Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \ .$$

In a sense $L_{\tilde{\boldsymbol{q}}}$ is the "best" classifier in the class of piecewise polynomial classifiers. It is well known that these

interpolating polynomials have good local approximation properties for Hölder smooth functions, namely we have that

Lemma 3.

$$\sup_{g^* \in \Sigma(L,\alpha)} \max_{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}} |L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})| = O(M^{-\alpha}) \ . \tag{20}$$

Lemma 3 is proved in the end of the section. We have almost all the pieces we need to conclude the proof. The last fact we need is a bound on the variation of the tensor-product Lagrange polynomials, namely it is easily shown that

$$\max_{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}} \left| Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \right| \leq \lfloor \alpha \rfloor^{(d-1)\lfloor \alpha \rfloor} \ . \tag{21}$$

We are now ready to show the final result. Assume for now that Ω_n holds, therefore $|\widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) - g^*(M^{-1}\tilde{\boldsymbol{l}})| \leq t_N$ for all $\tilde{\boldsymbol{l}}$. Note that t_N is decreasing as n (and consequently N) increase.

$$\begin{aligned}
R(\widehat{G}_n) - R(G^*) &= O\left(\|\widehat{g} - g^*\|_\kappa^\kappa \right) \\
&= O\left(\sum_{\tilde{\boldsymbol{q}} \in \{0,\ldots,M/\lfloor\alpha\rfloor - 1\}^{d-1}} \left\| (\widehat{L}_{\tilde{\boldsymbol{q}}} - g^*) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa^\kappa \right) \\
&= O\left(\sum_{\tilde{\boldsymbol{q}}} \left\| (L_{\tilde{\boldsymbol{q}}} - g^*) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} + (\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}}) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa^\kappa \right) \\
&= O\left(\sum_{\tilde{\boldsymbol{q}}} \left(\left\| (L_{\tilde{\boldsymbol{q}}} - g^*) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa + \left\| (\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}}) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa \right)^\kappa \right) \ ,
\end{aligned}$$

Note now that

$$\begin{aligned}
\left\| (L_{\tilde{\boldsymbol{q}}} - g^*) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa &= \left(\int_{I_{\tilde{\boldsymbol{q}}}} \left(L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}}) \right)^\kappa \mathrm{d}\tilde{\boldsymbol{x}} \right)^{1/\kappa} \\
&= O\left(\left(\int_{I_{\tilde{\boldsymbol{q}}}} M^{-\alpha\kappa} \mathrm{d}\tilde{\boldsymbol{x}} \right)^{1/\kappa} \right) = O\left(M^{-\alpha} M^{-\frac{d-1}{\kappa}} \right) \ .
\end{aligned}$$

Where we used Lemma 3. We have also

$$\begin{aligned}
\left\| (\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}}) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa &= \sum_{\tilde{\boldsymbol{l}} : M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} \left| \widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) - g^*(M^{-1}\tilde{\boldsymbol{l}}) \right| \left\| Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}} \right\|_\kappa \\
&\leq \sum_{\tilde{\boldsymbol{l}} : M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} t_N \left(\int_{I_{\tilde{\boldsymbol{q}}}} \left| Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \right|^\kappa \mathrm{d}\tilde{\boldsymbol{x}} \right)^{1/\kappa} \\
&\leq \sum_{\tilde{\boldsymbol{l}} : M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} t_N \left(\int_{I_{\tilde{\boldsymbol{q}}}} \lfloor \alpha \rfloor^{(d-1)\lfloor\alpha\rfloor\kappa} \mathrm{d}\tilde{\boldsymbol{x}} \right)^{1/\kappa} = O\left(t_N M^{-(d-1)/\kappa} \right) \ .
\end{aligned}$$

Using these two facts we conclude that

$$\begin{aligned} R(\widehat{G}_n) - R(G^*) = & \\ & O\left(\sum_{\tilde{\boldsymbol{q}}}\left(\|(L_{\tilde{\boldsymbol{q}}} - g^*)\mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}\|_\kappa + \left\|(\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}})\mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}\right\|_\kappa\right)^\kappa\right) \\ = & O\left(\sum_{\tilde{\boldsymbol{q}} \in \{0,\ldots,M/\lfloor\alpha\rfloor - 1\}^{d-1}}\left(M^{-\alpha}M^{-\frac{d-1}{\kappa}} + t_N M^{-(d-1)/\kappa}\right)^\kappa\right) \\ = & O\left(M^{d-1}\left(M^{-\alpha}M^{-\frac{d-1}{\kappa}} + t_N M^{-(d-1)/\kappa}\right)^\kappa\right) \\ = & O\left(\left(M^{-\alpha} + t_N\right)^\kappa\right) \ . \end{aligned}$$

Plugging in the choices of M and N given in the theorem statement we obtain

$$R(\widehat{G}_n) - R(G^*) = O\left((\log n/n)^{\frac{\alpha\kappa}{\alpha(2\kappa-2)+d-1}}\right) \ .$$

Finally, noticing that $1 - \Pr(\Omega_n) = O\left(n^{-\frac{\alpha\kappa}{\alpha(2\kappa-2)+d-1}}\right)$ we have

$$\begin{aligned} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) &\leq O\left((\log n/n)^{\frac{\alpha\kappa}{\alpha(2\kappa-2)+d-1}}\right)\Pr(\Omega_n) + 1\cdot(1-\Pr(\Omega_n)) \\ &= O\left((\log n/n)^{\frac{\alpha\kappa}{\alpha(2\kappa-2)+d-1}}\right) \ , \end{aligned}$$

concluding the proof.

■

A. *Proof of Lemma 3*

Let $\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}$ and $g \in \Sigma(L,\alpha)$. Taking into account Definition 1 we have

$$\begin{aligned} |L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})| &= \left|L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}}) + \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(\tilde{\boldsymbol{x}})\right| \\ &\leq \left|L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(\tilde{\boldsymbol{x}})\right| + \left|g^*(\tilde{\boldsymbol{x}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(\tilde{\boldsymbol{x}})\right| \\ &\leq \left|L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(\tilde{\boldsymbol{x}})\right| + L\left\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}\right\|^\alpha \\ &\leq \left|L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(\tilde{\boldsymbol{x}})\right| + O(M^{-\alpha}) \ . \end{aligned}$$

Note now that the tensor polynomial approximation space contains the space of degree $\lfloor \alpha \rfloor$ polynomials, therefore we can write $L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}})$ as a tensor product polynomial and so

$$
\begin{aligned}
|L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})| &\leq \left| \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} g^*(M^{-1}\tilde{\boldsymbol{l}}) Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(\tilde{\boldsymbol{x}}) \right| + O(M^{-\alpha}) \\
&= \left| \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} \left(g^*(M^{-1}\tilde{\boldsymbol{l}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(M^{-1}\tilde{\boldsymbol{l}}) \right) Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \right| + O(M^{-\alpha}) \\
&\leq \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} \left| g^*(M^{-1}\tilde{\boldsymbol{l}}) - \mathrm{TP}_{\tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1}}(M^{-1}\tilde{\boldsymbol{l}}) \right| \left| Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \right| + O(M^{-\alpha}) \\
&\leq \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} L \left\| \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1} \right\|^{\alpha} \left| Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \right| + O(M^{-\alpha}) \qquad (22) \\
&\leq \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} L \left\| \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{q}}\lfloor\alpha\rfloor M^{-1} \right\|^{\alpha} \lfloor\alpha\rfloor^{(d-1)\lfloor\alpha\rfloor} + O(M^{-\alpha}) \\
&\leq \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} O(M^{-\alpha}) + O(M^{-\alpha}) \\
&= \lfloor\alpha\rfloor^{d-1} O(M^{-\alpha}) + O(M^{-\alpha}) = O(M^{-\alpha}) \ ,
\end{aligned}
$$

where we applied Definition 1 again, and in step (22) we used (21). Finally the last step follows from the observation that the number of terms in the summation is $\lfloor\alpha\rfloor^{d-1}$, which does not depend on M.

■

REFERENCES

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

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

[3] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, no. 2-3, pp. 133–168, August 1997.

[4] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Learning probabilistic linear-threshold classifiers via selective sampling," in *The Sixteenth Annual Conference on Learning Theory. LNAI 2777, Springer*, 2003.

[5] G. Blanchard and D. Geman, "Hierarchical testing designs for pattern recognition," *The Annals of Statistics*, vol. 33, no. 3, pp. 1155–1202, 2005.

[6] S. Dasgupta, A. Kalai, and C. Monteleoni, "Analysis of perceptron-based active learning," in *Eighteen Annual Conference on Learning Theory (COLT)*, 2005.

[7] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Advances in Neural Information Processing (NIPS)*, 2005.

[8] ——, "Analysis of a greedy active learning strategy," in *Advances in Neural Information Processing (NIPS)*, 2004.

[9] N. Balcan, A. Beygelzimer, and J. Langford, "Agostic active learning," in *23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, 2006.

[10] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2005, extended version available at http://homepages.cae.wisc.edu/∼rcastro/ECE-05-3.pdf.

[11] M. Kääriäinen, "On active learning in the non-realizable case," NIPS Workshop on Foundations of Active Learning, 2005.

[12] M. Horstein, "Sequential decoding using noiseless feedback," *IEEE Trans. Info. Theory*, vol. 9, no. 3, pp. 136–143, 1963.

[13] M. V. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problems in Information Transmission*, vol. 10, pp. 223–231, 1974, (Translated from *Problemy Peredachi Informatsii,* 10(3):51–61, July-September, 1974. Original article submitted June 25, 1973).

[14] P. Hall and I. Molchanov, "Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces," *The Annals of Statistics*, vol. 31, no. 3, pp. 921–941, 2003.

[15] G. Golubev and B. Levit, "Sequential recovery of analytic periodic edges in the binary image models," *Mathematical Methods of Statistics*, vol. 12, pp. 95–115, 2003.

[16] B. Bryan, J. Schneider, R. C. Nichol, C. J. Miller, C. R. Genovese, and L. Wasserman, "Active learning for identifying function threshold boundaries," in *Advances in Neural Information Processing (NIPS)*, 2005.

[17] A. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *The Annals of Statistics*, vol. 32, no. 1, pp. 135–166, 2004.

[18] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, pp. 171–186, 2005.

[19] S. K. Thompson and G. A. F. Seber, *Adaptive Sampling*. New York: John Wiley & Sons, Inc., 1996.

[20] M. Ghosh, N. Mukhopadhyay, and P. Sen, *Sequential Estimation*. John Wiley & Sons, Inc., 1997.

[21] P. K. Sen, *Sequential Nonparametrics*. John Wiley & Sons, Inc., 1981.

[22] E. Baum, "Neural net algorithms that learn is polynomial time from examples and queries," *IEEE Transaction on Neural Networks*, vol. 2, pp. 5–19, 1991.

[23] L. Cavalier, "Nonparametric estimation of regression level sets," *Statistics*, vol. 29, pp. 131–160, 1997.

[24] A. B. Tsybakov, "On nonparametric estimation of density level sets," *Annals of Statistics*, vol. 25, pp. 948–969, 1997.

[25] ——, *Introduction à l'estimation non-paramétrique*, ser. Mathématiques et Applications, 41. Springer, 2004.

[26] R. Castro and R. Nowak, "Upper and lower bounds for active learning," in *44th Annual Allerton Conference on Communication, Control and Computing*, 2006.

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

[28] D. Donoho, "Wedgelets: Nearly minimax estimation of edges," *The Annals of Statistics*, vol. 27, pp. 859–897, 1999.

[29] A. P. Korostelev, "On minimax rates of convergence in image models under sequential design," *Statistics & Probability Letters*, vol. 43, pp. 369–375, 1999.

[30] A. Korostelev and J.-C. Kim, "Rates of convergence for the sup-norm risk in image models under sequential designs," *Statistics & probability Letters*, vol. 46, pp. 391–399, 2000.

[31] C. de Boor, "The error in polynomial tensor-product, and Chung-Yao, interpolation," in *Surface Fitting and Multiresolution Methods*, A. LeMéhauté, C. Rabut, and L. Schumaker, Eds. Vanderbilt University Press, 1997, pp. 35–50.