Minimax Bounds for Active Learning

Rui M. Castro[1,2] and Robert D. Nowak[1]

[1] University of Wisconsin, Madison WI 53706, USA,
rcastro@cae.wisc.edu,nowak@engr.wisc.edu,
[2] Rice University, Houston TX 77005, USA

Abstract. 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.

1 Introduction

The interest in active learning in the machine learning community has increased greatly in the last few of years, in part due to the dramatic growth of data sets and the high cost of labeling all the examples in such sets. There are several empirical and theoretical results suggesting that in certain situations active learning can be significantly more effective than passive learning [1–5]. Many of these results pertain to the "noiseless" setting, in which the labels are deterministic functions of the features. 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 need 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, 6–8]. 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 active learning results have been extended to the "bounded noise rate" setting. In this setting labels are no longer a deterministic function of the features, but for a given feature the probability of observing a particular label is significantly higher than the probability of observing 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 $|\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) - 1/2| > c$ 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 to active learning procedures. Under this assumption it can be shown that results similar to the ones for the noiseless scenario can be achieved

[4, 9–11]. These results are intimately related to adaptive sampling techniques in regression problems [12–14, 10, 15], where similar performance gains have been reported. Furthermore the active learning algorithm proposed in [9] in addition to provide improvements in certain bounded noise conditions is shown to perform no worse than passive learning in general settings.

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 procedures can follow, since as sampling approaches the decision boundary the conditional probability $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ approaches $1/2$. Since situations like this seem very likely to arise in practice (e.g., simply due to feature measurement 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 [16], the complexity of the Bayes decision boundary can in many cases be characterized by a parameter $\rho = (d-1)/\alpha$, where d is the dimension of the feature space and α is the Hölder regularity of the boundary. Furthermore, the behavior of $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ in the vicinity of the boundary can be characterized by a parameter $\kappa \geq 1$. The value $\kappa = 1$ corresponds to the noiseless or bounded noise situation and $\kappa > 1$ corresponds to unbounded noise conditions. We derive lower bounds on 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 collect examples, whereas the fastest decay rate possible using passive learning is $n^{-\frac{\kappa}{2\kappa+\rho-1}}$. Note that the active learning rate is always superior to that of passive learning. Tsybakov has shown that in certain cases ($\kappa \to 1$ 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 efficiently drive an active learning procedure. Furthermore we 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_X (over features) is bounded from above and below on $[0, 1]^d$.

The paper is organized as follows. In Section 2 we formally state the active learning problem and define the probability classes under consideration. Section 3 presents the basic results on lower bounds for active learning rates and in Section 4 we provide corresponding upper bounds, which match the lower bounds up to a logarithmic factor. Together, this demonstrates the bounds are tight and hence near minimax optimal. Final remarks are made in Section 5 and the main proofs are given in the Appendix.

2 Problem Formulation

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

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

R^* is attained by 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 *conditional probability* (we use this term only if it is clear from the context). In general $R(G^*) > 0$ 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*

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

where Δ denotes the symmetric difference between two sets[3], and $P_{\boldsymbol{X}}$ is the marginal distribution of $\boldsymbol{X}$.

Suppose that $P_{\boldsymbol{X}Y}$ is unknown, but that we have a large (infinite) pool of feature examples we can select 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 operation 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, where one queries the label of a feature vector. Formally we have:

A1 - Y_i, $i \in \{1, \ldots, n\}$ are distributed as

$$Y_i = \begin{cases} 1 \ , \text{ with probability } \eta(\boldsymbol{X}_i) \\ 0 \ , \text{ with probability } 1 - \eta(\boldsymbol{X}_i) \end{cases} .$$

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

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

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

$$\boldsymbol{X}_i|\boldsymbol{X}_1\dots\boldsymbol{X}_{i-1},\boldsymbol{X}_{i+1},\dots,\boldsymbol{X}_n,Y_1\dots Y_{i-1},Y_{i+1},\dots,Y_n \overset{\text{a.s.}}{=} \boldsymbol{X}_i|\boldsymbol{X}_1\dots\boldsymbol{X}_{i-1},Y_1\dots Y_{i-1}\ .$$

The conditional distribution on the right hand side (r.h.s) of the above expression is called the *sampling strategy* and is denoted by S_n. It completely defines our sampling procedure. After collecting the n examples, that is after collecting $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$, we construct a classifier $\widehat{G}_n$ that is desired to be close to G^*. 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 $\boldsymbol{X}_j$ and Y_i), and therefore the collection of sample points $\{\boldsymbol{X}_i\}_{i=1}^n$ may be chosen before any observations 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).

In this paper we are interested in a particular class of distributions, namely scenarios where the Bayes decision set is a boundary fragment. That is, the Bayes decision boundary is the graph of function. We consider Hölder smooth boundary functions. Throughout the paper assume that $d \geq 2$, the dimension of the feature space.

Definition 1. *A function $f : [0,1]^{d-1} \to \mathbb{R}$ is* Hölder smooth *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 $L, \alpha > 0$, and $TP_{\boldsymbol{x}}(\cdot)$ denotes the order 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 $\text{epi}(g) = \left\{(\boldsymbol{x}, y) \in [0,1]^{d-1} \times [0,1] : y \geq g(\boldsymbol{x})\right\}$, that is, $\text{epi}(g)$ is epigraph of g. Define

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

In other words $\mathcal{G}_{\text{BF}}$ is a collection of sets indexed by Hölder smooth functions of the first $d-1$ coordinates of the feature domain $[0,1]^d$. Therefore G^* and the corresponding boundary function g^* are equivalent representations of the Bayes classifier.

In order to get a better understanding of the potential of active learning we impose further conditions on the distribution $P_{\boldsymbol{X}Y}$. We assume that $P_{\boldsymbol{X}}$ is uniform on $[0,1]^d$. The results in this paper can easily be generalized to the case where the marginal density of $\boldsymbol{X}$ with respect to the Lebesgue measure is not uniform, but bounded above and below, yielding the same rates of error

convergence. 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

$$|\eta(\boldsymbol{x}) - 1/2| \geq c|x_d - g^*(\tilde{\boldsymbol{x}})|^{\kappa-1}, \quad \text{if } |x_d - g^*(\tilde{\boldsymbol{x}})| \leq \epsilon_0 \ , \tag{2}$$

$$|\eta(\boldsymbol{x}) - 1/2| \geq c{\epsilon_0}^{\kappa-1}, \quad \text{if } |x_d - g^*(\tilde{\boldsymbol{x}})| > \epsilon_0 \ , \tag{3}$$

for some $\epsilon_0 > 0$. The condition above is very similar to the so-called margin condition (or noise-condition) introduced by Tsybakov [16]. If $\kappa = 1$ then the $\eta(\cdot)$ function "jumps" across the Bayes decision boundary, that is $\eta(\cdot)$ is bounded away from the value $1/2$. If $\kappa > 1$ then $\eta(\cdot)$ crosses the value $1/2$ at the Bayes decision boundary. Condition (2) indicates that $\eta(\cdot)$ cannot be arbitrarily "flat" around the decision boundary (*e.g.*, for $\kappa = 2$ the function $\eta(\cdot)$ behaves linearly around $1/2$). This means that the noise affecting observations that are made close to the decision boundary is roughly proportional to the distance to the boundary. We also assume a reverse-sided condition on $\eta(\cdot)$, namely

$$|\eta(\boldsymbol{x}) - 1/2| \leq C|x_d - g^*(\tilde{\boldsymbol{x}})|^{\kappa-1} \ , \tag{4}$$

for all $\boldsymbol{x} \in [0,1]^d$, where $C > c$. This condition, together with (2) and (3) provides a two-sided characterization of the "noise" around the decision boundary. Similar two-sided conditions have been proposed for other problems [17, 18]. Let $\mathrm{BF}(\alpha, \kappa, L, C, c)$ be the class of distributions satisfying the noise conditions above with parameter κ and whose Bayes classifiers are boundary fragments with smoothness α.

3 Lower Bounds

In this section we present lower bounds on the performance of active and passive sampling methods. We start by characterizing active learning for the boundary fragment classes.

Theorem 1. *Let $\rho = (d-1)/\alpha$. Then*

$$\liminf_{n\to\infty} \inf_{\widehat{G}_n, S_n} \sup_{P \in BF(\alpha,\kappa,L,C,c)} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) \geq c_{\min} n^{-\frac{\kappa}{2\kappa+\rho-2}} \ ,$$

where $\inf_{\widehat{G}_n, S_n}$ denotes the infimum over all possible classifiers and sampling strategies S_n, and $c_{\min} > 0$ is a constant.

The proof of Theorem 1 is presented in Appendix A. An important remark is that condition (4) does not play a role in the rate of the lower bound, therefore dropping that assumption (equivalently taking $C = \infty$) does not alter the result of the theorem.

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 interval $[0,1]^d$ in a somewhat uniform manner. These can be deterministically placed, for example over a uniform grid, or simply taken

uniformly distributed over $[0,1]$. The results in [16] imply that, under (A1), (A2.1), and $\kappa \geq 1$,

$$\inf_{\widehat{G}_n, S_n} \sup_{P \in \mathrm{BF}(\alpha,\kappa,L,C,c)} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) \geq c_{\min} n^{-\frac{\kappa}{2\kappa+\rho-1}} \tag{5}$$

where the samples $\{\boldsymbol{X}_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) uniformly over $[0,1]^d$. Furthermore this 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). The active learning lower bound is also tight (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 1 is a lower bound, and it therefore applies to the broader classes of distributions introduced in [16], characterized in terms of the metric entropy of the class of Bayes classifiers.

The proof of Theorem 1 employs relatively standard techniques, and follows the approach in [19]. 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 managing assumption (A2.2) are the key aspects of the proof. Notice also that the result in (5) can be obtained by modifying the proof of Theorem 1 slightly.

4 Upper Bounds

In this section we construct an active learning procedure and upper bound its error performance. The upper bound achieves the rates of Theorem 1 to within a logarithmic factor. This procedure yields a classifier $\widehat{G}_n$ that has boundary fragment structure, although the boundary is no longer a smooth function. It is instead a piecewise polynomial function. This methodology proceeds along the lines of [20, 21], extending one-dimensional active sampling methods to this higher dimensional setting. For this methodology we use some results reported in [22] addressing the problem of one-dimensional change-point detection under the noise conditions imposed in this paper. The ideas in that work were motivated by the work of Burnashev and Zigangirov [12], pertaining a change-point detection problem under the bounded noise rate condition (equivalent to $\kappa = 1$).

We begin by constructing a grid over the first $d-1$ dimensions of the feature domain, namely let M be an integer and $\tilde{\boldsymbol{l}} \in \{0, \ldots, M\}^{d-1}$. Define the line segments $\mathcal{L}_{\tilde{\boldsymbol{l}}} \stackrel{\Delta}{=} \{(M^{-1}\tilde{\boldsymbol{l}}, x_d) : x_d \in [0,1]\}$. We collect N samples along each line, yielding a total of NM^{d-1} samples (where $n \geq NM^{d-1}$). Our goal is to estimate $g(M^{-1}\tilde{\boldsymbol{l}})$, for all $\tilde{\boldsymbol{l}}$, using these samples. We will then interpolate the estimates of g at these points to construct a final estimate of the decision boundary. The correct choices for M and N will arise from the performance analysis; for now

we point out only that both M and N are growing with the total number of samples n.

When restricting ourselves to the line segment $\mathcal{L}_{\tilde{\boldsymbol{l}}}$ the estimation problem boils down to a one-dimensional change-point detection problem. Consider first the case $\kappa = 1$. In [12] an active sampling methodology was developed and analyzed, with the following property: using N sample points actively chosen yields an estimator $\widehat{g}(M^{-1}\tilde{\boldsymbol{l}})$ of $g(M^{-1}\tilde{\boldsymbol{l}})$ such that

$$\Pr\left(|\widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) - g^*(M^{-1}\tilde{\boldsymbol{l}})| > t\right) \leq \frac{1}{t}\exp(-c^2 N) \ ,$$

therefore it is possible to estimate $g^*(M^{-1}\tilde{\boldsymbol{l}})$ accurately with a very small number of samples. It was shown in [22] (and further detailed in [23]) that, when $\kappa > 1$, using N sample points in $\mathcal{L}_{\tilde{\boldsymbol{l}}}$ chosen actively based on knowledge of κ, yields an estimate $\widehat{g}(M^{-1}\tilde{\boldsymbol{l}})$ of $g(M^{-1}\tilde{\boldsymbol{l}})$ such that

$$\Pr(|\widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) - g^*(M^{-1}\tilde{\boldsymbol{l}})| > t) \leq \frac{2}{t}\exp\left(-\frac{N}{3}c^2\left(\frac{t}{6}\right)^{2\kappa-2}\right) \ . \tag{6}$$

Taking

$$t = t_N \stackrel{\Delta}{=} c_1 \left(\log N/N\right)^{\frac{1}{2\kappa-2}} \tag{7}$$

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

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. Begin by dividing $[0,1]^{d-1}$ (that is, the domain of g^*) into cells. Let M_0 be the largest integer such that $M_0 \leq M/\lfloor\alpha\rfloor$. Let $\tilde{\boldsymbol{q}} \in \{0,\ldots,M_0\}^{d-1}$ index the cells

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

Note that these cells almost partition the domain $[0,1]^{d-1}$ entirely. If $M/\lfloor\alpha\rfloor$ is not an integer there is a small region on the edge of the domain that is not covered by these cells, with volume $O(M^{-1})$. In each of these cells we perform a polynomial interpolation using the estimates of g^* at points within the cell. 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 $\tilde{\boldsymbol{x}} \in [0,1]^d$. The functions $Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}$ are the tensor-product Lagrange polynomials [24]. The final estimate of g^* is therefore given by

$$\widehat{g}(\tilde{\boldsymbol{x}}) = \sum_{\tilde{\boldsymbol{q}} \in \{0,\ldots,M_0\}^{d-1}} \widehat{L}_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}$$

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

Theorem 2. *Consider the classification methodology described above, using* $M = \left\lfloor n^{\frac{1}{\alpha(2\kappa-2)+d-1}} \right\rfloor$ *and* $N = \lfloor n/(M-1)^{d-1} \rfloor$. *Let* $\rho = (d-1)/\alpha$, *then*

$$\limsup_{n\to\infty} \sup_{P\in BF(\alpha,\kappa,L,C,c)} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) \leq c_{\max} \left(\log n/n\right)^{\frac{\kappa}{2\kappa+\rho-2}} \quad .$$

The proof of Theorem 2 is given in Appendix B. One sees that this estimator achieves the rate of Theorem 1 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 [20] that if $\kappa, \alpha = 1$ the logarithmic factor can be eliminated by using a slightly more sophisticated interpolation scheme.

5 Final Remarks

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 error decay rates. Our upper bounds were derived constructively, based on an active learning procedure originally developed for one-dimensional change-point detection [12]. 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). A potential approach is a multiscale technique as used in [10]. The results of this paper do indicate what we should be aiming for in terms of performance. Moreover, the bounds clarify the situations in which active learning may or may not offer a significant gain 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.

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

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 large class $\mathrm{BF}(\alpha, \kappa, L, C, c)$ to a test of a finite set of hypothesis. These are distributions $P_{\boldsymbol{X}Y} \in \mathrm{BF}(\alpha, \kappa, L, C, c)$ chosen carefully. The main tool is the following theorem, adapted from [19] (page 85, theorem 2.5).

Theorem 3 (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* $d_f(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ *be a collection*

of semi-distances (indexed by $f \in \mathcal{F}$). Suppose we have $\{f_0, \ldots, f_M\} \in \mathcal{F}$ such that

i) $d_{f_k}(f_j, f_k) \geq 2\,a > 0, \quad \forall_{0 \leq j,k \leq M}$,
ii) $P_{f_0} \ll P_{f_j}, \quad \forall_{j=1,\ldots,M}$, *(see footnote[4])*
iii) $\frac{1}{M} \sum_{j=1}^{M} KL(P_{f_j} \| P_{f_0}) \leq \gamma \log M$,

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

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} P_f \left(d_f(\widehat{f}, f) \geq a \right) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\gamma - 2\sqrt{\frac{\gamma}{\log M}} \right) > 0 \ ,$$

where the infimum is taken with respect to the collection of all possible estimators of f (based on a sample from P_f), and KL denotes the Kullback-Leibler divergence[5] .

Note that in final statement of Theorem 3 the semi-distance between the estimate $\widehat{f}_n$ and $\hat{f}$ might depend on f. This is a critical feature in our setup, since the excess risk depends on the underlying unknown distribution (1).

To apply the theorem we need to construct a subset of $\mathrm{BF}(\alpha, \kappa, L, C, c)$ with the desired characteristics. These elements 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 $\tilde{\boldsymbol{l}} \in \{1, \ldots, m\}^{d-1}$. Define also $\varphi_{\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) = L m^{-\alpha} h(m(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}))$, 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 $[0,1]^{d-1}$.

---

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

[5] Let P and Q be two probability measures defined on a common probability space. 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.

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

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

The set $\mathcal{G}$ is a collection of 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 $\mathcal{G} \subseteq \Sigma(L, \alpha)$. Let $g \in \mathcal{G}$ and construct the conditional distribution

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

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

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 (2), (3) and (4) are satisfied for the distributions above. By construction the Bayes decision boundary for each of these distributions is given by $x_d = g(\tilde{\boldsymbol{x}})$ 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 3. Recall the following lemma.

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

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

where ρ denotes the Hamming distance.

For a proof of the Lemma 1 see [19](page 89, lemma 2.7). To apply Theorem 3 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^*. Define the semidistances

$$d_i(G, G') = \int_{G \Delta G'} |2\eta_{\boldsymbol{\omega}^{(i)}}(\boldsymbol{x}) - 1| \mathrm{d}\boldsymbol{x} .$$

The next step of the proof is to lower-bound $d_i(G_j^*, G_i^*) = R_i(G_j^*) - R_i(G_i^*)$ for all $j \ne i$. Note that

$$\begin{aligned} d_i(G_j^*, G_i^*) &= \int_{[0,1]^{d-1}} \int_0^{|g_i^*(\tilde{\boldsymbol{x}}) - g_j^*(\tilde{\boldsymbol{x}})|} |2\eta_{\boldsymbol{\omega}^{(i)}}(\boldsymbol{x}) - 1| \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\ &= \sum_{\tilde{\boldsymbol{l}} \in \{1,\ldots,m\}^{d-1}} |\boldsymbol{\omega}_{\tilde{\boldsymbol{l}}}^{(i)} - \boldsymbol{\omega}_{\tilde{\boldsymbol{l}}}^{(j)}| \int_{[0,1]^{d-1}} \int_0^{L m^{-\alpha} h(m(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}))} |2\eta_{\boldsymbol{\omega}^{(i)}}(\boldsymbol{x}) - 1| \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} . \end{aligned}$$

To bound the double-integral we just need to consider the two possible values of $\boldsymbol{\omega}_{\tilde{\boldsymbol{l}}}^{(i)}$. We display here case $\boldsymbol{\omega}_{\tilde{\boldsymbol{l}}}^{(i)} = 1$, but exactly the same result can be shown for $\boldsymbol{\omega}_{\tilde{\boldsymbol{l}}}^{(i)} = 0$.

$$
\begin{aligned}
&\int_{[0,1]^{d-1}} \int_0^{Lm^{-\alpha}h(m(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}))} |2\eta_{\boldsymbol{\omega}^{(i)}}(\boldsymbol{x}) - 1| \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\
&= \int_{[0,1]^{d-1}} \int_0^{Lm^{-\alpha}h(m(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}}))} 2c(x_d - Lm^{-\alpha}h(m(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_{\tilde{\boldsymbol{l}}})))^{\kappa-1} \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{x}} \\
&= 2cm^{-(d-1)} \int_{[-1/2,1/2]^{d-1}} \int_0^{Lm^{-\alpha}h(\tilde{\boldsymbol{z}})} (x_d - Lm^{-\alpha}h(\tilde{\boldsymbol{z}})^{\kappa-1} \mathrm{d}x_d \mathrm{d}\tilde{\boldsymbol{z}} \\
&= \frac{2cm^{-(d-1)}}{\kappa} \int_{[-1/2,1/2]^{d-1}} L^\kappa m^{-\alpha\kappa} h^\kappa(\tilde{\boldsymbol{z}}) \mathrm{d}\tilde{\boldsymbol{z}} \\
&= \frac{2c}{\kappa} L^\kappa m^{-\alpha\kappa-(d-1)} \|h\|_\kappa^\kappa \sim m^{-\alpha\kappa-(d-1)} \ ,
\end{aligned}
$$

where $\|h\|_\kappa$ denotes the κ norm of h. Taking into account Lemma 1 we have that, for n large enough

$$
\begin{aligned}
d_i(G_j^*, G_i^*) &\geq \rho(\boldsymbol{\omega}_{\tilde{\boldsymbol{l}}}^{(i)}, \boldsymbol{\omega}_{\tilde{\boldsymbol{l}}}^{(j)}) \frac{2c}{\kappa} L^\kappa m^{-\alpha\kappa-(d-1)} \|h\|_\kappa^\kappa \\
&\geq \frac{2c}{8\kappa} L^\kappa \|h\|_\kappa^\kappa \, m^{-\alpha\kappa} \triangleq a_n \sim m^{-\alpha\kappa} \ .
\end{aligned}
$$

We are ready for the final step of the proof. We need the following straightforward result

Lemma 2. *Let P and Q be Bernoulli random variables with parameters respectively p and q, such that $p, q \to 1/2$. Then $KL(P\|Q) = 2(p-q)^2 + o((p-q)^2)$.*

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 $\eta_{\boldsymbol{\omega}^{(i)}}$. Use the notation $\boldsymbol{Z}_j^X \triangleq (\boldsymbol{X_1}, \ldots, \boldsymbol{X_j})$ and $\boldsymbol{Z}_j^Y \triangleq (Y_1, \ldots, Y_j)$. Then

$$
\begin{aligned}
&\mathrm{KL}(P_i\|P_0) = \mathbb{E}_i \left[\log \frac{P_{\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y;i}(\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y)}{P_{\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y;0}(\boldsymbol{Z}_n^X, \boldsymbol{Z}_n^Y)} \right] \\
&= \mathbb{E}_i \left[\log \frac{\prod_{j=1}^n P_{Y_j|\boldsymbol{X_j};i}(Y_j|\boldsymbol{X_j}) \ P_{\boldsymbol{X_j}|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y;i}(\boldsymbol{X_j}|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y)}{\prod_{j=1}^n P_{Y_j|\boldsymbol{X_j};0}(Y_j|\boldsymbol{X_j}) \ P_{\boldsymbol{X_j}|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y;0}(\boldsymbol{X_j}|\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y)} \right] \qquad (8) \\
&= \mathbb{E}_j \left[\log \frac{\prod_{j=1}^n P_{Y_j|\boldsymbol{X}_j;i}(Y_j|\boldsymbol{X}_j)}{\prod_{j=1}^n P_{Y_j|\boldsymbol{X}_j;0}(Y_j|\boldsymbol{X}_j)} \right] \\
&= \mathbb{E}_j \left[\mathbb{E}_j \left[\log \frac{\prod_{j=1}^n P_{Y_j|\boldsymbol{X}_j;i}(Y_j|\boldsymbol{X}_j)}{\prod_{j=1}^n P_{Y_j|\boldsymbol{X}_j;0}(Y_j|\boldsymbol{X}_j)} \middle| \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \right] \right] \\
&\leq 2n(cA^{\kappa-1})^2 + o(n(cA^{\kappa-1})^2) \leq \mathrm{const} \cdot nm^{-\alpha(2\kappa-2)} \ ,
\end{aligned}
$$

where the last inequality holds provided n is large enough and const is chosen appropriately. In (8) note that the distribution of $\boldsymbol{X_j}$ conditional on $\boldsymbol{Z}_{j-1}^X, \boldsymbol{Z}_{j-1}^Y$ depends only on the sampling strategy S_n, and therefore does not change with the underlying distribution, hence those terms in the numerator and denominator cancel out. Finally

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

From Lemma 1 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 3 and so

$$\inf_{\widehat{G}_n, S_n} \sup_{P \in \mathrm{BF}(\alpha,\kappa,L,C,c)} P(R(\widehat{G}_n) - R(G^*) \geq a_n) \geq c_{\min} \ ,$$

where $c_{\min} > 0$, for n large enough. An application of Markov's inequality yields the original statement of the theorem, concluding the proof.

B Proof of Theorem 2

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' [6] notation for simplicity. 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 (6) and the choice t_N in (7) 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 (7) 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

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

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

where the inequality follows from condition (4).

Let $L_{\tilde{\boldsymbol{q}}}$, $\tilde{\boldsymbol{q}} \in \{0, \ldots, M_0\}^{d-1}$ be the clairvoyant version of $\widehat{L}_{\tilde{\boldsymbol{q}}}$, that is,

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

It is well known that these interpolating polynomials have good local approximation properties for Hölder smooth functions, namely we have that

$$\sup_{g \in \Sigma(L,\alpha)} \max_{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}} |L_{\tilde{\boldsymbol{q}}}(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})| = O(M^{-\alpha}) \ . \tag{9}$$

This result is proved in [23]. 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

$$\max_{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}} \left| Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}}) \right| \leq \lfloor \alpha \rfloor^{(d-1)\lfloor \alpha \rfloor} \ . \tag{10}$$

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.

$$\begin{aligned}
&R(\widehat{G}_n) - R(G^*) = O\left(\|\widehat{g} - g^*\|_\kappa^\kappa\right) \\
&= O\left(\sum_{\tilde{\boldsymbol{q}} \in \{0,\ldots,M_0\}^{d-1}} \left\| (\widehat{L}_{\tilde{\boldsymbol{q}}} - g^*) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa^\kappa \right) + O(M^{-1}) \\
&= O\left(\sum_{\tilde{\boldsymbol{q}}} \left\| (L_{\tilde{\boldsymbol{q}}} - g^*) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} + (\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}}) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa^\kappa \right) + O(M^{-1}) \\
&= O\left(\sum_{\tilde{\boldsymbol{q}}} \left(\|(L_{\tilde{\boldsymbol{q}}} - g^*) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}\|_\kappa + \left\| (\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}}) \mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\} \right\|_\kappa \right)^\kappa \right) + O(M^{-1}) \ ,
\end{aligned}$$

where the term $O(M^{-1})$ corresponds to the error in the area around the edge of $[0,1]^{d-1}$, not covered by any cells in $\{I_{\tilde{\boldsymbol{q}}}\}$. The volume of this region is $O(M^{-1})$. Note now that

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

Where we used (9). We have also

$$\left\|(\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}})\mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}\right\|_\kappa = \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} \left|\widehat{g}(M^{-1}\tilde{\boldsymbol{l}}) - g^*(M^{-1}\tilde{\boldsymbol{l}})\right| \left\|Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}\right\|_\kappa$$

$$\leq \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} t_N \left(\int_{I_{\tilde{\boldsymbol{q}}}} \left|Q_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{l}}}(\tilde{\boldsymbol{x}})\right|^\kappa \mathrm{d}\tilde{\boldsymbol{x}}\right)^{1/\kappa}$$

$$\leq \sum_{\tilde{\boldsymbol{l}}: M^{-1}\tilde{\boldsymbol{l}} \in I_{\tilde{\boldsymbol{q}}}} t_N \left(\int_{I_{\tilde{\boldsymbol{q}}}} \lfloor\alpha\rfloor^{(d-1)\lfloor\alpha\rfloor\kappa} \mathrm{d}\tilde{\boldsymbol{x}}\right)^{1/\kappa} = O\left(t_N M^{-(d-1)/\kappa}\right) .$$

Using these two facts we conclude that

$$\begin{aligned}
&R(\widehat{G}_n) - R(G^*) = \\
&\quad O\left(\sum_{\tilde{\boldsymbol{q}}} \left(\|(L_{\tilde{\boldsymbol{q}}} - g^*)\mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}\|_\kappa + \left\|(\widehat{L}_{\tilde{\boldsymbol{q}}} - L_{\tilde{\boldsymbol{q}}})\mathbf{1}\{\tilde{\boldsymbol{x}} \in I_{\tilde{\boldsymbol{q}}}\}\right\|_\kappa\right)^\kappa\right) + O(M^{-1}) \\
&= O\left(\sum_{\tilde{\boldsymbol{q}} \in \{0,\ldots,M_0\}^{d-1}} \left(M^{-\alpha} M^{-\frac{d-1}{\kappa}} + t_N M^{-(d-1)/\kappa}\right)^\kappa\right) + O(M^{-1}) \\
&= O\left(M^{d-1}\left(M^{-\alpha} M^{-\frac{d-1}{\kappa}} + t_N M^{-(d-1)/\kappa}\right)^\kappa\right) + O(M^{-1}) \\
&= O\left(\left(M^{-\alpha} + t_N\right)^\kappa + M^{-1}\right) .
\end{aligned}$$

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.

References

1. Mackay, D.J.C.: Information-based objective functions for active data selection. Neural Computation **4** (1991) 698–714
2. Cohn, D., Ghahramani, Z., Jordan, M.: Active learning with statistical models. Journal of Artificial Intelligence Research (1996) 129–145
3. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning **28**(2-3) (August 1997) 133–168

4. Cesa-Bianchi, N., Conconi, A., Gentile, C.: Learning probabilistic linear-threshold classifiers via selective sampling. In: The Sixteenth Annual Conference on Learning Theory. LNAI 2777, Springer. (2003)
5. Blanchard, G., Geman, D.: Hierarchical testing designs for pattern recognition. The Annals of Statistics **33**(3) (2005) 1155–1202
6. Dasgupta, S., Kalai, A., Monteleoni, C.: Analysis of perceptron-based active learning. In: Eighteen Annual Conference on Learning Theory (COLT). (2005)
7. Dasgupta, S.: Coarse sample complexity bounds for active learning. In: Advances in Neural Information Processing (NIPS). (2005)
8. Dasgupta, S.: Analysis of a greedy active learning strategy. In: Advances in Neural Information Processing (NIPS). (2004)
9. Balcan, N., Beygelzimer, A., Langford, J.: Agostic active learning. In: 23rd International Conference on Machine Learning, Pittsburgh, PA, USA (2006)
10. Castro, R., Willett, R., Nowak, R.: 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. Kääriäinen, M.: On active learning in the non-realizable case. NIPS Workshop on Foundations of Active Learning (2005)
12. Burnashev, M.V., Zigangirov, K.S.: An interval estimation problem for controlled observations. Problems in Information Transmission **10** (1974) 223–231 (Translated from *Problemy Peredachi Informatsii,* 10(3):51–61, July-September, 1974. Original article submitted June 25, 1973).
13. Hall, P., Molchanov, I.: Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. The Annals of Statistics **31**(3) (2003) 921–941
14. Golubev, G., Levit, B.: Sequential recovery of analytic periodic edges in the binary image models. Mathematical Methods of Statistics **12** (2003) 95–115
15. Bryan, B., Schneider, J., Nichol, R.C., Miller, C.J., Genovese, C.R., Wasserman, L.: Active learning for identifying function threshold boundaries. In: Advances in Neural Information Processing (NIPS). (2005)
16. Tsybakov, A.: Optimal aggregation of classifiers in statistical learning. The Annals of Statistics **32**(1) (2004) 135–166
17. Cavalier, L.: Nonparametric estimation of regression level sets. Statistics **29** (1997) 131–160
18. Tsybakov, A.B.: On nonparametric estimation of density level sets. Annals of Statistics **25** (1997) 948–969
19. Tsybakov, A.B.: Introduction à l'estimation non-paramétrique. Mathématiques et Applications, 41. Springer (2004)
20. Korostelev, A.P.: On minimax rates of convergence in image models under sequential design. Statistics & Probability Letters **43** (1999) 369–375
21. Korostelev, A., Kim, J.C.: Rates of convergence for the sup-norm risk in image models under sequential designs. Statistics & probability Letters **46** (2000) 391–399
22. Castro, R., Nowak, R.: Upper and lower bounds for active learning. In: 44th Annual Allerton Conference on Communication, Control and Computing. (2006)
23. Castro, R.M., Nowak, R.D.: Minimax bounds for active learning. Technical report, ECE Dept., University of Wisconsin - Madison (2007) (available at http://homepages.cae.wisc.edu/∼rcastro/ECE-07-3.pdf).
24. de Boor, C.: The error in polynomial tensor-product and chung-yao, interpolation. In LeMéhauté, A., Rabut, C., Schumaker, L., eds.: Surface Fitting and Multiresolution Methods, Vanderbilt University Press (1997) 35–50