An Adaptive Strategy for Active Learning with Smooth Decision Boundary

Andrea Locatelli ANDREA.LOCATELLI@OVGU.DE
Department of Mathematics
Otto-von-Güricke Universität Magdeburg

Alexandra Carpentier ALEXANDRA.CARPENTIER@OVGU.DE
Department of Mathematics
Otto-von-Güricke Universität Magdeburg

Samory Kpotufe SAMORY@PRINCETON.EDU
Operations Research and Financial Engineering
Princeton University

Abstract

We present the first adaptive strategy for active learning in the setting of classification with smooth decision boundary. The problem of adaptivity (to unknown distributional parameters) has remained opened since the seminal work of Castro and Nowak (2007), which first established (active learning) rates for this setting. While some recent advances on this problem establish *adaptive* rates in the case of univariate data, adaptivity in the more practical setting of multivariate data has so far remained elusive. Combining insights from various recent works, we show that, for the multivariate case, a careful reduction to univariate-adaptive strategies yield near-optimal rates without prior knowledge of distributional parameters.

1. Introduction

In active learning (for classification), the learner can *actively* request Y labels at any point x in the data space to speedup learning: the goal is to return a classifier with low error while requesting as few labels as possible. Previous work (see e.g. Freund et al. (1993); Castro and Nowak (2007); Hanneke (2009); Koltchinskii (2010); Minsker (2012); Balcan et al. (2009)) showed that under various distributional settings, active learning offers a significant advantage over passive learning (the usual classification setting with i.i.d. labeled data).

An important such setting is the one studied in the seminal work of Castro and Nowak (2007), known as the boundary fragment setting, where the feature space $\mathcal{X} = [0,1]^d$ is bisected along the d-th coordinate by a smooth curve which characterizes the decision boundary $\{x : \mathbb{E}[Y|x] = 1/2\}$. The essential error measure in this setting is the distance from the estimated decision boundary to the true decision boundary; such error metric can readily serve to bound the usual 0-1 classification error under additional distributional assumptions, e.g.,assuming that the marginal P_X is uniform as done in Castro and Nowak (2007) (we will relax such assumptions). They show that the minimax optimal rate (in terms of excess 0-1 error over the Bayes classifier) achievable by an active strategy is strictly faster than in the passive setting of Tsybakov (2004). While their strategy is minimax optimal, it is unfortunately non-adaptive, i.e., it requires full knowledge of key distributional parameters. Namely, there are two important such parameters: α, which captures the *smoothness* of the decision boundary, and κ, which controls the *noise rate*, i.e. *how fast* $\mathbb{E}[Y|x]$ grows away from $1/2$ near
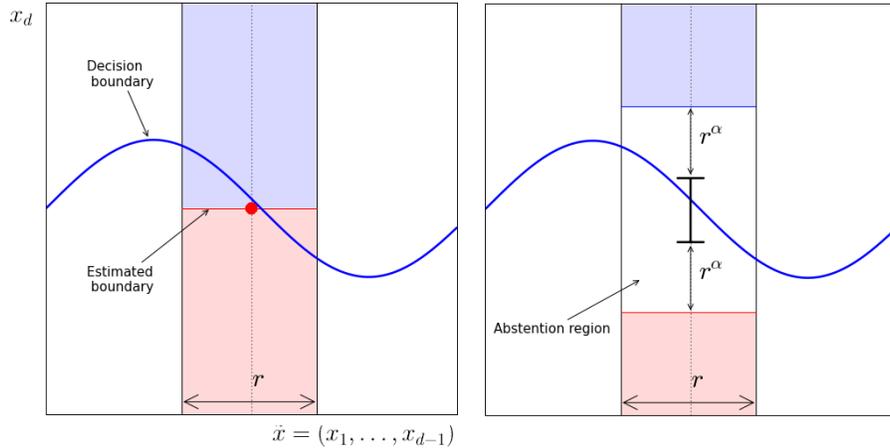
Figure 1: Comparing strategies (for known $\alpha \leq 1$). On the left, the strategy in Castro and Nowak (2007); Yan et al. (2016). Both strategies operate on fixed grids with cells of side-length r, and perform a line search in each cell (dotted line). A threshold (the red dot) guaranteed to be close to the decision boundary is returned, and then extrapolated to the entire cell (estimated boundary). The strategy needs to operate on an optimal value of $r = r(\alpha, \kappa)$. On the right (our strategy), the line search returns an interval of size $O(r^\alpha)$, guaranteed to intersect the decision boundary; the interval is then extended by $O(r^\alpha)$ to create an *abstention* region of the right size (in terms of known α). To adapt to unknown κ, the strategy is repeated over dyadic values of $r \to 0$.

the decision boundary. These parameters interpolate between hard and easy problems (rough or smooth decision boundary, high or low noise), and are never known in practice. Therefore, a minimax adaptive strategy – i.e., one which attains optimal rates but does not require a priori knowledge of such parameters – is highly desirable. Such optimal adaptive strategy has unfortunately remained elusive for the general case of data in $\mathbb{R}^d$.

For univariate data ($d = 1$), it is known (Hanneke (2009); Ramdas and Singh (2013)) that this limitation can be overcome, and minimax optimal strategies (such as the A^2 algorithm in Balcan et al. (2009), further studied in Hanneke (2007)) exist, which adapt to unknown noise rate κ on the line (there is no notion of smoothness α in the line setting since the boundary is just a threshold). Recently, earlier results of Koltchinskii (2010); Hanneke et al. (2011) – meant for settings with *bounded disagreement coefficients* – were extended in Wang (2011) to obtain an adaptive procedure for the boundary fragment class of Castro and Nowak (2007), including the case of data in $\mathbb{R}^d$; unfortunately that strategy yields suboptimal rates for the setting.

We present the first adaptive and optimal strategy for the setting, by combining insights from various recent work on related problems, and original insights from Castro and Nowak (2007).

Combining insights from related work. The original strategy of Castro (2007) consists of a clever reduction of active learning in $\mathbb{R}^d$ to active learning on $\mathbb{R}$: since the boundary is the curve of function $g : [0,1]^{d-1} \mapsto [0,1]$, (a) first partition $[0,1]^{d-1}$ into a finite number of cells, and do active learning on each cell as follows: (b) pick a line on the cell, and estimate the threshold at which the decision boundary crosses this line; (c) extrapolate the estimated threshold to the whole cell using the fact that the boundary is smooth. Unfortunately step (a) required knowledge of both κ and α to pick an optimal cell size, while steps (b) and (c) respectively required knowledge of noise margin κ and smoothness α. This strategy is illustrated in Figure 1 (left box).

A key step in our work, is to temporarily assume knowledge of α and to aim for a procedure that is adaptive to κ, while following the above strategy of Castro and Nowak (2007). Clearly, given recent advances on adaptive active learning on $\mathbb{R}$, step (a) above is readily made adaptive to κ. This

2

is for instance done in the recent work of Yan et al. (2016), which however leaves open the problem in (a) of choosing a partition of optimal cell-size in terms of unknown κ (their work and this issue is discussed in more detail in Section 4). We show that we can resolve this issue by proceeding hierarchically over decreasing cell sizes. Furthermore, in order to eventually adapt to unknown α, we also require a small but crucial change to the interpolation in step (c) above (the same essential interpolation strategy is used in both Castro and Nowak (2007); Yan et al. (2016). The reason for more careful interpolation is described next.

In order to adapt to unknown smoothness α, we build on recent insights from Locatelli et al. (2017) which concerns a separate classification setting with smooth regression function $\eta(x) \doteq \mathbb{E}[Y|x]$ rather than smooth decision boundary. Their work presents a generic adaptive strategy that exploits the nested structure of smoothness classes, namely the fact that an α-smooth function is also α'-smooth for any $\alpha' < \alpha$. Their strategy consists of aggregating the classification estimates returned by a subroutine taking increasing smoothness values α' as a parameter. The subroutine in our case is that described in the last paragraph – which takes in the smoothness as a parameter. As it turns out, for the aggregation to work, the subroutine has to be *correct* in a sense that is suitable to our setting, namely, for any $\alpha' < \alpha$, it must only label points that are at an optimal distance away from the decision boundary and *abstain* otherwise (see Figure 1). In other words, the interpolation step (c) discussed above, must produce an *abstention* region of optimal radii in terms of κ and α.

Thus, the bulk of our analysis is in constructing a sub-procedure that takes in α as a parameter, is fully adaptive to κ, and properly abstains in regions of optimal size in terms of α and unknown κ. Our construction readapts the line-search in Yan et al. (2016) to our particular needs and constraints.

2. Setting

In this section, we describe formally the problem of active learning under nonparametric assumptions in the membership query setting.

2.1 The Active Learning Setting

Binary Classification. We write $\mathbb{P}_{X,Y}$ for the joint-distribution of feature-label pairs (X, Y). $\mathbb{P}_X$ denotes the marginal distribution according to variable X, supported on $[0,1]^d$. The random variable Y belongs to $\{0, 1\}$ as usual in the binary classification setting. The conditional distribution of Y knowing $X = x$, which we denote $\mathbb{P}_{Y|X=x}$, is characterized by the regression function

$$\eta(x) \doteq \mathbb{E}[Y|X=x], \quad \forall x \in [0,1]^d.$$

The Bayes classifier is defined as $f^*(x) = \mathbf{1}\{\eta(x) \geq 1/2\}$. It minimizes the 0-1 risk $R(f) = \mathbb{P}_{X,Y}(Y \neq f(X))$ over all possible $f : [0,1]^d \mapsto \{0, 1\}$. The aim of the learner is to return a classifier f with small excess error

$$\mathcal{E}(f) \doteq R(f) - R(f^*) = \int_{x \in [0,1]^d : f(x) \neq f^*(x)} |1 - 2\eta(x)| \mathrm{d}\mathbb{P}_X(x). \tag{1}$$

Active sampling. At each time $t \leq n$, the active learner can sample a label Y at any $x_t \in [0,1]^d$ drawn from the conditional distribution $\mathbb{P}_{Y|X=x_t}$. In total, it can sample at most $n \in \mathbb{N}^*$ labels - we will refer to n as the *sampling budget* - known to the learner. At the end of the budget, the active learner returns a classifier $\widehat{f}_n : [0,1]^d \mapsto \{0, 1\}$.

In this work, our goal is to design an adaptive sampling strategy that outputs a good estimate of the decision boundary, with high probability over the samples requested and labels revealed, without prior knowledge of distributional parameters, i.e., smoothness and noise margin parameters. This is formalized in Section 2.2 below.

2.2 The Nonparametric Setting

In this section, we expose our assumptions on $\mathbb{P}_{X,Y}$, which are nonparametric in nature, and similar to the setting introduced in Castro and Nowak (2007). From now on, we assume that $d \geq 2$.

Definition 1 (Hölder smoothness). *We say that a function $g : [0,1]^{d-1} \mapsto [0,1]$ belongs to the Hölder class $\Sigma(\lambda, \alpha)$ if g is $\lfloor \alpha \rfloor^1$ times continuously differentiable and for all $x, y \in [0,1]^{d-1}$, and any $\beta \leq \alpha$ we have:*

$$|g(x) - \mathrm{TP}_{y,\lfloor \beta \rfloor}(x)| \leq \lambda \|x - y\|_\infty^\beta, \tag{2}$$

where $\mathrm{TP}_{y,\lfloor \beta \rfloor}$ is the Taylor polynomial expansion of degree $\lfloor \beta \rfloor$ of g in y and $\|z\|_\infty \doteq \max_{1 \leq i \leq d} |z_i|$ is the usual infinity norm for d dimensional vectors.

For any $g \in \Sigma(\lambda, \alpha)$, consider the set $\mathrm{epi}(g) \doteq \{x = (\tilde{x}, x_d) \in [0,1]^{d-1} \times [0,1] : x_d \geq g(\tilde{x})\}$, which is the epigraph of the function g. We define the boundary fragment class $\mathcal{G}(\lambda, \alpha) \doteq \{\mathrm{epi}(g), g \in \Sigma(\lambda, \alpha)\}$.

Assumption 1 (Smoothness of the boundary). *There exists constants $\alpha > 0$ and $\lambda \geq 1$ such that $\{x : \eta(x) \geq 1/2\} \in \mathcal{G}(\lambda, \alpha)$.*

In other words, there exists $g^* \in \Sigma(\lambda, \alpha)$ such that $\{x : \eta(x) \geq 1/2\} = \mathbb{1}\{\mathrm{epi}(g^*)\}$ and the Bayes classifier is equivalent to $\mathbb{1}\{\mathrm{epi}(g^*)\}$. This means that the decision boundary for the classification problem is fully characterized by $g^* \in \Sigma(\lambda, \alpha)$. Importantly, for any $\alpha' \leq \alpha$, we also have $g^* \in \Sigma(\lambda, \alpha')$, as the classes $\Sigma(\lambda, \alpha) \subset \Sigma(\lambda, \alpha')$ are nested for λ fixed.

We also assume a one-sided noise condition on the behavior of the regression function close to the decision boundary characterized by $g^* \in \Sigma(\lambda, \alpha)$, which can be seen as a geometric variant of the popular Tsybakov noise condition (TNC)(Tsybakov (2004)).

Assumption 2 (Geometric TNC). *There exists constants $c > 0$ and $\kappa \geq 1$ such that for any $x = (\tilde{x}, x_d) \in [0,1]^{d-1} \times [0,1]$:*

$$|\eta(x) - \frac{1}{2}| \geq c|x_d - g^*(\tilde{x})|^{\kappa - 1}.$$

This assumption characterizes how "flat" the regression function η is allowed to be in the vicinity of the decision boundary: the larger κ the noise parameter, the harder it is to locate the decision boundary precisely. In particular, for $\kappa = 1$, η "jumps" at the decision boundary, going from $1/2 - c$ to $1/2 + c$.

In this work, our main objective is to devise an adaptive algorithm that returns an estimate $\hat{g}$ of the true decision boundary g^*, such that $\|\hat{g} - g^*\|_\infty$ is small and of optimal size in a minimax sense. Under additional assumptions (which relax original assumptions in Castro (2007)), we will show that the resulting classifier $x = (\tilde{x}, x_d) \to \mathbf{1}\{x_d \geq \hat{g}(\tilde{x})\}$ also attains optimal excess risk guarantees.

Definition 2. *We denote $\mathcal{P}(\alpha, \kappa) \doteq \mathcal{P}(\lambda, \alpha, \kappa, c)$ the set of classification problems $\mathbb{P}_{X,Y}$ characterized by $(\mathbb{P}_X, \eta)$ such that Assumption 1 is satisfied for some $g^* \in \Sigma(\lambda, \alpha)$ and Assumption 2 is satisfied with constants $\kappa \geq 1, c > 0$.*

For the rest of the paper we will consider $c > 0$ to be fixed, and $\lambda \geq 1$ to be fixed and known to the learner - we discuss the relevance of this assumption in Section 3.1. Now, considering κ to be fixed as well as λ, we remark that the nested structure of the smoothness classes straightforwardly implies the same property for the classes $\mathcal{P}(\alpha, \kappa)$.

---

1. $\lfloor \alpha \rfloor$ denotes the largest integer strictly smaller than α.

3. Main Results

In this section, we show our mains results, assuming we have access to a black-box Subroutine with some correctness property. We first formalize this notion of correctness, and deduct from this a property of the aggregation procedure, which allows us to then state our main adaptive results.

3.1 Adaptive Algorithm

A first component of our adaptive strategy is a meta-procedure (Algorithm 1) that aggregates the classification estimates of a subroutine that takes α as a parameter (but must adapt to unknown noise margin κ). While much of our analysis concerns this Subroutine, this section introduces the meta-procedure whose definition is needed for stating the main result of Theorem 1.

The metaprocedure implements original ideas from the recent work of Locatelli et al. (2017) (which itself adapts ideas in Lepski and Spokoiny (1997) to the active setting), which considers a different distributional setting (smoothness of η rather than smoothness of the boundary g^*) but with a similar nested structure as in this work. The conditions on the Subroutine for the meta-procedure to work in our setting are different, as we will see, and designing a suitable such subroutine constitutes the bulk of our efforts.

Algorithm 1 Adapting to unknown boundary smoothness α

Input: n, δ, λ, and a black-box Subroutine
Initialization: $s_0^0 = s_0^1 = \emptyset$
for $i = 1, ..., \lfloor \log(n) \rfloor^2$ **do**
    Let $n_0 = \frac{n}{\lfloor \log(n) \rfloor^2}$, $\delta_0 = \frac{\delta}{\lfloor \log(n) \rfloor^2}$, and $\alpha_i = \frac{i}{\lfloor \log(n) \rfloor}$
    Run Subroutine with parameters $(n_0, \delta_0, \alpha_i, \lambda)$ and receive S_i^0, S_i^1
    For $y \in \{0, 1\}$, set $s_i^y = s_{i-1}^y \cup (S_i^y \setminus s_{i-1}^{1-y})$
end for
Output:

- Confidently labeled sets $S^0 = s_{\lfloor \log(n) \rfloor^2}^0, S^1 = s_{\lfloor \log(n) \rfloor^2}^1$,
- Estimated Boundary: $\hat{g}_n(\tilde{x}) \doteq \min\{x_d : (\tilde{x}, x_d) \in S_1\}$
- Classifier $\hat{f}_n(x) \doteq \mathbf{1}\{x \in S^1\} = \mathbf{1}\{x_d \geq \hat{g}(\tilde{x})\}$

The subroutine is called over increasing guesses α_i of the unknown smoothness parameter α of the boundary, taking advantage of the nested nature of the Hölder classes: if g^* is α-Hölder for some unknown α, then it is α_i-Hölder for $\alpha_i \leq \alpha$. Crucially, the subroutine labels only part of the space, and *abstains* otherwise. Now, suppose that the subroutine, called on α_i, guarantees *correctly* labeled sets S_i^0, S_i^1 whenever g^* is α_i-Hölder; then for any $\alpha_i \leq \alpha$ the aggregated labels remain correct. When $\alpha_i > \alpha$, the Subroutine might return incorrect labels. However, this is not a problem since the aggregation procedure never overwrites previously assigned labels, and thus misclassification only occurs in the *abstention* region returned by previous calls with $\alpha_i \leq \alpha$. Thus, as long as these abstention regions are of optimal size w.r.t. $\alpha_i \leq \alpha$, the final error of the aggregation procedure will be of optimal order (provided some $\alpha_i \approx \alpha$).

Following the above intuition, we now formally define *correctness* in a sense suited to our particular setting and implicit goal of estimating the decision boundary. This is different from the notion of *correctness* in Locatelli et al. (2017) where the goal is to achieve a correct margin Δ w.r.t. the regression function η, i.e. finding x s.t. $|\eta(x) - 1/2| > \Delta$, rather than finding x that are Δ distant from the boundary $\{x : \eta(x) = 1/2\}$ as in our case.

Definition 3 ((δ, Δ, n)-correct algorithm)**.** *Consider a procedure which returns disjoint measurable sets $S^0, S^1 \subset [0,1]^d$. Let $0 < \delta < 1$, and $\Delta \geq 0$. We call such a procedure **weakly** (δ, Δ, n)-**correct** for a classification problem $\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \kappa)$ if, with probability larger than $1 - 2\delta$ using at most n*

label requests:

$$\left\{ x = (\tilde{x}, x_d) \in [0,1]^d : x_d - g^*(\tilde{x}) > \Delta \right\} \subset S^1$$
$$\left\{ x = (\tilde{x}, x_d) \in [0,1]^d : g^*(\tilde{x}) - x_d > \Delta \right\} \subset S^0.$$

If in addition, under the same probability event over at most n label requests, we have

$$S^1 \subset \left\{ x = (\tilde{x}, x_d) \in [0,1]^d : x_d > g^*(\tilde{x}) \right\}$$
$$S^0 \subset \left\{ x = (\tilde{x}, x_d) \in [0,1]^d : x_d < g^*(\tilde{x}) \right\}$$

then such a procedure is simply called (δ, Δ, n)*-***correct** *for* $\mathbb{P}_{X,Y}$.

In the boundary fragment setting, correctness is defined in terms of distance to the decision boundary, which is a major difference with respect to the smooth regression function (see Locatelli et al. (2017) and the different notion of correctness therein). Importantly, a correct procedure returns labeled sets with the following key properties (with high probability): first, points are always labeled in agreement with their true class (and thus, bring no excess risk). Second, it abstains in a region of width at most Δ around the true decision boundary.

3.2 Main Results

In this Section we present our main result, Theorem 1, which bounds the distance from our estimated boundary to the true boundary. As a corollary, the excess 0-1 risk of the estimated classifier can be bounded under additional distributional assumptions that relax the original setting of Castro (2007).

We start with the following simple proposition, stating (as in the intuition detailed above) that Algorithm 1 correctly aggregates estimates whenever the subroutine calls return correct estimates.

Proposition 1 (Correctness of aggregation)**.** *Let* $n \in \mathbb{N}^*$ *and* $1 > \delta > 0$. *Let* $\delta_0 = \delta/(\lfloor \log(n) \rfloor^2)$ *and* $n_0 = n/(\lfloor \log(n) \rfloor^2)$ *as in Algorithm 1. Fix* $\kappa \geq 1$. *Suppose that, for any* $\alpha > 0$, *the Subroutine in Algorithm 1 is* $(\delta_0, \Delta_\alpha, n_0)$*-correct for any* $\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \kappa)$, *where* $\Delta_\alpha > 0$ *depends on* n, δ *and the class* $\mathcal{P}(\alpha, \kappa)$.

Fix $\alpha \in [\lfloor \log(n) \rfloor^{-1}, \lfloor \log(n) \rfloor]$, *and let* $\alpha_i = i/\lfloor \log(n) \rfloor$ *for* $i \in \{1, \ldots, \lfloor \log(n) \rfloor^2\}$. *Then Algorithm 1 is* **weakly** $(\delta_0, \Delta_{\alpha_i}, n_0)$*-correct for any* $\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \kappa)$ *for the largest* i *such that* $\alpha_i \leq \alpha$.

The proof of this proposition follows can be found in Section A.3 of the Appendix, and follows from arguments in Locatelli et al. (2017). The main difference in the interpretation of this result with respect to the result in Locatelli et al. (2017), in which correctness is defined in terms of distance between η and $1/2$, is that we are interested here in locating the decision boundary g^*. This makes Proposition 1 very simple to visualize in our setting. For any run with $\alpha_i \leq \alpha$, the decision boundary is estimated within a margin Δ_{α_i} such that no regions are misclassified. As $\alpha_i \leq \alpha$ grows, this margin decreases, until it reaches the largest i^* such that $\alpha_{i^*} \leq \alpha$. For any $i > i^*$, we cannot characterize the behavior of the non-adaptive Subroutine; fortunately, the misclassified regions are confined to the set $\left\{ x = (\tilde{x}, x_d) \in [0,1]^d : |x_d - g^*(x)| < \Delta_{\alpha_{i^*}} \right\}$.

We now state our main adaptive result (Theorem 1). Following Proposition 1, the main work in obtaining Theorem 1 consist of producing a Subroutine that is *correct* in the sense of Definition 3. This is done in Theorem 4 of Section 4.

Theorem 1. *Let* $n \in \mathbb{N}^*$ *and* $\delta > 0$. *Assume that* $\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \kappa)$ *with* $\alpha \in [\lfloor \log(n) \rfloor^{-1}, \lfloor \log(n) \rfloor]$. *Algorithm 1 run with parameters* (n, δ, λ) *and using Algorithm 2 as the black-box Subroutine outputs an approximation of the decision boundary* $\hat{g}_n$ *such that with probability at least* $1 - 2\delta$:

$$||\hat{g}_n - g^*||_\infty \leq C \lambda^{\frac{(d-1)}{2\alpha(\kappa-1)+d-1}} \left(\frac{\log^3(n/\delta)}{n} \right)^{\alpha/(2\alpha(\kappa-1)+d-1)},$$

where $C > 0$ *is a constant that does not depend on* λ, n, δ.

The proof of this Theorem can be found in Section A.3. By setting $\delta = n^{-\log(n)/(d-1)}$ in Theorem 1, we also get a rate in expectation of order $\tilde{O}\left(n^{-\alpha/(2\alpha(\kappa-1)+d-1)}\right)$, matching (up to logarithmic factors) the minimax lower bound derived in Castro and Nowak (2007).

So far, we have made no assumption on $\mathbb{P}_X$. In order to relate this bound on the distance between $\hat{g}_n$ and g^* to a guarantee on the risk of the classifier $\hat{f}_n$, we now state a third assumption, which bounds the risk incurred by regions that are close to g^*.

Assumption 3. *There exists $C > 0$, $\Delta_0 > 0$ and $\kappa' > 0$ such that $\forall \Delta \in [0, \Delta_0]$:*

$$\int_{x \in [0,1]^d : |x_d - g^*(\tilde{x})| \leq \Delta} |1 - 2\eta(x)| \mathrm{d}\mathbb{P}_X(x) \leq C\Delta^{\kappa'}$$

This assumption relaxes the setting introduced in Castro and Nowak (2007), as we will see in Example 1 (in particular there, $\kappa' = \kappa$ which can be strong). Assumption 3 and Theorem 1 directly lead to the following corollary, which bounds the excess risk of the classifier with high probability.

Corollary 1. *Under the assumptions of Theorem 1 and Assumption 3, for $n \geq N = N(\alpha, \lambda, \kappa, \delta)$, Algorithm 1 run with (n, δ, λ) outputs a classifier $\hat{f}_n$ such that with probability at least $1 - 2\delta$ its excess risk is bounded as:*

$$\mathcal{E}(\hat{f}_n) \leq C\lambda^{\frac{\kappa'(d-1)}{2\alpha(\kappa-1)+d-1}} \left(\frac{\log^3(n/\delta)}{n}\right)^{\alpha\kappa'/(2\alpha(\kappa-1)+d-1)} \quad ,$$

where $C > 0$ is a constant that does not depend on λ, n, δ.

From the corollary we see that larger values of κ' and lower values for κ improve the rate; this can be a source of tension under the restriction that $\kappa' = \kappa$ as in the first example below. The first example below is the exact setting of Castro and Nowak (2007).

Example 1 (Castro and Nowak (2007)). Consider $\mathbb{P}_X$ uniform over $[0, 1]^d$ and η such that:

$$c|x_d - g^*(\tilde{x})|^{\kappa-1} \leq \left|\eta(x) - \frac{1}{2}\right| \leq C|x_d - g^*(\tilde{x})|^{\kappa-1}.$$

It is clear that Assumption 3 is satisfied with $\kappa' = \kappa$. Under these assumptions, the minimax rate in expectation for the excess risk is of order $\Omega(n^{-\alpha\kappa/(2\alpha(\kappa-1)+d-1)})$ as shown by Castro and Nowak (2007). Our procedure is the first adaptive and optimal (up to logarithmic factors) strategy in this setting. Notice that in this case, both low and large values of κ seem to improve the rate. In fact, for $\alpha > (d-1)/2$ we get *fast rates* (below $n^{-1/2}$) and lower values of κ improve the rate. On other hand, when $\alpha \leq (d-1)/2$, greater values of κ improve the rate. This tension comes from the fact that lower values of κ on the one hand make it easier to locate the decision boundary as there is a sharper jump close to g^*; yet for large values of $\kappa = \kappa'$, misclassifying a large region close to the boundary bears less risk. Assumption 3 decouples the effect of κ and κ', which is evident in the following example.

Example 2 (Hard and soft margin in $\mathbb{P}_X$). Consider situations where $\mathbb{P}_X$ has little or no mass near the decision boundary. First consider the extreme of no mass near the boundary (hards margin), i.e. there exists Δ_0 such that

$$\mathbb{P}_X(x : |x_d - g^*(\tilde{x})| \leq \Delta_0) = 0.$$

In this case $\kappa' = \infty$ in Assumption 3, and the classifier attains 0 error with high probability (equivalently, exponentially small error in expectation). More generally (soft-margin), Assumption 3 holds if $\mathbb{P}_X$ decreases sufficiently fast near the boundary: for instance, suppose we have $\forall 0 < \Delta \leq \Delta_0$,

$$\mathbb{P}_X(x : |x_d - g^*(\tilde{x})| \leq \Delta) \leq \Delta^{\kappa'-\kappa_0+1},$$

7

where $\kappa_0 \le \kappa \wedge (\kappa' + 1)$ satisfies the upper-bound $\left|\eta(x) - \frac{1}{2}\right| \le c|x_d - g^*(\tilde{x})|^{\kappa_0 - 1}$.

We complete this result with the following lower bound, which shows that the rate in Corollary 1 is tight up to logarithmic factors, at least for $\kappa' > \kappa - 1$, and strictly faster than the passive rate under the same assumptions.

Theorem 2 (Active Lower Bound). *Let $\alpha > 0, \kappa > 1$ and $\kappa' > \kappa - 1$. Consider $\mathcal{P}(\alpha, \kappa, \kappa')$ the subset of $\mathcal{P}(\alpha, \kappa)$ such that Assumption 3 is satisfied with κ'. For n large enough, any (possibly active) strategy $\mathcal{A}_n$ that collects at most n samples before returning a classifier $\widehat{f}_n$ satisfies:*

$$\inf_{\mathcal{A}_n} \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \kappa, \kappa')} \mathbb{E}[\mathcal{E}(\widehat{f}_n)] \ge Cn^{-\alpha\kappa'/(2\alpha(\kappa-1)+d-1)},$$

where $C > 0$ does not depend on n and the expectation is taken with respect to both the samples collected by the strategy $\mathcal{A}_n$ and $\mathbb{P}_{X,Y}$.

Finally, we derive a lower bound in the passive setting, in terms of κ' (previous lower-bounds for related settings do not consider κ', see for example Tsybakov (2004)). The lower-bound below highlights the gains in active learning, as the rate of Corollary 1 and Theorem 2 is strictly faster than the passive-learning lower-bound obtained below.

Theorem 3 (Passive Lower Bound). *Under the Assumption of Theorem 2, for n large enough, any classifier $\widehat{f}_n$ trained on at most n i.i.d. samples satisfies:*

$$\inf_{\widehat{f}_n} \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \kappa, \kappa')} \mathbb{E}[\mathcal{E}(\widehat{f}_n)] \ge Cn^{-\alpha\kappa'/(\alpha(\kappa+\kappa'-1)+d-1)},$$

where $C > 0$ does not depend on n.

For $\kappa = \kappa'$, the lower-bound recovers known rates for passive learning see e.g. Tsybakov (2004); Castro (2007).

The proofs of these Theorems can be found in the Section A.4 and A.5 of the Appendix. It is based on general information theoretic arguments (Fano's method) as exposed in a suitable form by Tsybakov (2009) and adapted to active learning by Castro and Nowak (2007). The geometric construction builds on lower-bound constructions in Locatelli et al. (2017) for the separate setting of smooth regression functions.

4. Analysis

4.1 A κ-Adaptive Procedure for the Boundary Fragment Class

We now introduce an algorithm that is fully adaptive with respect to κ the noise parameter, and takes as input α, λ the smoothness parameters of the decision boundary such that $g^* \in \Sigma(\lambda, \alpha)$. The strategy uses as a subroutine another adaptive procedure that solves the unidimensional problem of finding a threshold x_d^* such that for $\tilde{x} \in [0,1]^{d-1}$ fixed $g^*(\tilde{x}) = x_d^*$, we will refer to this univariate problem as the *line-search* problem in our context. In this section, we assume that we have access to a line-search procedure such that when it is called with a certain confidence δ and precision ϵ, it returns a threshold estimate T such that $|T - x_d^*| \le \epsilon$ with probability at least $1 - \delta$ using at most $\tilde{\mathcal{O}}(\epsilon^{-2(\kappa-1)})$ samples[2]. Such a procedure was proposed in the recent work of Yan et al. (2016). In their work, they use this procedure as a subroutine in the setting where one wants to estimate the boundary with a $\hat{g}$ such that $||\hat{g} - g^*|| \le \epsilon$ with high probability. Assuming knowledge of the smoothness α and given a target error ϵ, they can guarantee a number $n = n(\epsilon)$ of label requests optimal and adaptive in terms of unknown κ. Interestingly, given the goal of fixed target error ϵ, the problem of adaptive cell size as exposed in Section 1 seems to disappear: it's sufficient to partition

---

2. we use $\tilde{\mathcal{O}}$ to hide logarithmic factors in $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$

$[0, 1]^{d-1}$ into cells of size $\epsilon^{1/\alpha}$. The procedure they use is the same as the one exposed in Castro and Nowak (2007), as both strategies rely on a discretization of $[0, 1]^{d-1}$, launch a number of line-searches on a grid that covers the feature space, and then use the threshold estimates on this grid to construct a smooth approximation of the boundary such that $||g^* - g^*|| \leq \epsilon$. However, in our setting (and that of Castro (2007)) we instead fix a labeling budget n and aim to achieve an error $\epsilon = \epsilon(n)$ adaptive to unknown κ; in other words, to use the algorithmic strategy of Yan et al. (2016) we need knowledge of the optimal ϵ (which depend on unknown κ) in order to define an optimal partition cell size. Indeed, in this fixed budget setting, the strategy in Castro and Nowak (2007) uses both κ and α to find the right step-size for the discretization which is of order $\lfloor n^{-\alpha/(2\alpha(\kappa-1)+d-1)} \rfloor$. Our strategy bypasses this issue by proceeding hierarchically over a dyadic partition of $[0, 1]^{d-1}$. Our stopping criterion for the line-search procedure only depends on α, λ and the cell size, and allows our procedure to fully adapt to κ. As a last step, we carefully select the regions to label – and hence the abstention region – so as to make the procedure *correct* in the sense of Definition 3.

Algorithm 2 κ-adaptive procedure in d-dimension

Input: n, δ, λ, α
Initialisation: $l = 1$, $t = 0$
while $\{t < n\}$ **do**
    $M_l = \max(1, \lfloor \alpha \rfloor) 2^l$
    $\epsilon_l = \lambda 2^{-l\alpha}$
    $\delta_l = \delta(\max(1, \lfloor \alpha \rfloor) 2^{l(d+1)})^{-1}$
    for each $\tilde{a}$ in $\{0, ..., M_l\}^{d-1}$ **do**
        Run Subroutine 3 on the line $\mathcal{L}_{\tilde{a}}$ with parameters ϵ_l, δ_l
        Receive threshold estimate $T_{l,\tilde{a}}$ and budget used $N_{l,\tilde{a}}$
    end for
    Compute total budget used at depth l: $N_l = \sum_{\tilde{a}} N_{l,\tilde{a}}$
    $t = t + N_l$
    $l = l + 1$
end while
$l^* \doteq l - 1$ (final completed depth)
Fit $\lfloor \alpha \rfloor$-degree tensor-product Lagrange polynomial approximation of boundary using $(T_{l^*, \tilde{a}})_{\tilde{a}}$
$b_{l^*} \doteq \lambda \lceil \alpha \rceil^{d \lceil \alpha \rceil} M_{l^*}^{-\alpha}$ (bias term)
$S^0 \doteq \{x : x_d \leq \widehat{P}(\tilde{x}) - 4b_{l^*}\}$
$S^1 \doteq \{x : x_d \geq \widehat{P}(\tilde{x}) + 4b_{l^*}\}$
Output: S^y for $y \in \{0, 1\}$

Our procedure, Algorithm 2, takes as input n the maximum sampling budget, δ a confidence parameter, as well as λ and α the smoothness parameters such that $g^* \in \Sigma(\lambda, \alpha)$. While we assume here that λ is known to the learner, it is sufficient to call the procedure with a known upper-bound on the true parameter. This follows from the nested nature of the smoothness we consider here, as we have $\Sigma(\lambda, \alpha) \subset \Sigma(\lambda', \alpha)$ for any $\lambda' \geq \lambda$. Our analysis reveals that λ only has a multiplicative effect on the rate, therefore setting $\lambda \doteq \log(n)$ for n large enough only worsens the rate by a logarithmic factor. At each depth l, the algorithm launches $(M_l + 1)^{d-1}$ line-searches with $M_l = \max(1, \lfloor \alpha \rfloor) 2^l$, on a grid of step M_l^{-1}. Precisely, for each $\tilde{a} \in \{0, ..., M_l\}^{d-1}$ it launches a line-search instance using Algorithm 3 on the line segment $\mathcal{L}_{\tilde{a}} \doteq \{(M_l^{-1} \tilde{a}, x_d), x_d \in [0, 1]\}$ with confidence parameter $\delta_l = \delta(\max(1, \lfloor \alpha \rfloor) 2^{l(d+1)})^{-1}$ and precision $\epsilon_l = \lambda 2^{-l\alpha}$. Importantly, the precision with which the line-search procedure is called depends only on the step-size of the grid and the smoothness parameters λ and α, and not on κ. Heuristically, the precision of the line-search need not be greater than the precision of the nonparametric approximation of degree $\lfloor \alpha \rfloor$ of the boundary fit with the estimated thresholds on the grid of step size M_l^{-1}, which motivates our choice for ϵ_l. After each run indexed by $\tilde{a}$, it receives the estimated threshold $T_{l,\tilde{a}}$ and the budget used $N_{l,\tilde{a}}$. While the total budget used is less than the maximum allowed budget n, the discretization is refined and line-searches are initialized

with a higher precision parameter. Once the budget has run out, we use the estimated thresholds $(T_{l^*,\tilde{a}})_{\tilde{a}}$ at the last depth l^* such that all the line-searches have terminated to construct a polynomial interpolation of degree $\lfloor \alpha \rfloor$ of the boundary, as in the original strategy of Castro and Nowak (2007). In the case of $\alpha \leq 1$, we simply use in each cell a constant approximation that takes the value of the estimates $(T_{l^*,\tilde{a}})_{\tilde{a}}$, the details of which can be found in the proof of Theorem 4. In what follows, we assume $\alpha > 1$ and describe the approximation method for higher order smoothness.

To that effect, we will use the tensor-product Lagrange polynomials as in Castro and Nowak (2007) on slightly larger cells, to ensure that the number of estimated thresholds (coming form the line-searches) in those cells is enough to fit a $\lfloor \alpha \rfloor$-degree polynomial approximation. Let $\tilde{q} \in \{0, ..., \frac{M_{l^*}}{\lfloor \alpha \rfloor} - 1\}^{d-1}$ index the cells:

$$I_{\tilde{q}} \doteq \left[\tilde{q}_1 \lfloor \alpha \rfloor M_{l^*}^{-1}, (\tilde{q}_1 + 1)\lfloor \alpha \rfloor M_{l^*}^{-1} \right] \times ... \times \left[\tilde{q}_{d-1} \lfloor \alpha \rfloor M_{l^*}^{-1}, (\tilde{q}_{d-1} + 1)\lfloor \alpha \rfloor M_{l^*}^{-1} \right].$$

These cells partition $[0,1]^{d-1}$ entirely, as we have $M_{l^*} = \lfloor \alpha \rfloor 2^{l^*}$. We use the tensor-product Lagrange polynomial basis as in Castro and Nowak (2007), defined as follows:

$$Q_{\tilde{q},\tilde{a}}(\tilde{x}) \doteq \prod_{i=1}^{d-1} \prod_{\substack{0 \leq j \leq \lfloor \alpha \rfloor \\ j \neq \tilde{a}_i - \lfloor \alpha \rfloor \tilde{q}_i}} \frac{\tilde{x}_i - M_{l^*}^{-1}(\lfloor \alpha \rfloor \tilde{q}_i + j)}{M_{l^*}^{-1}\tilde{a}_i - M_{l^*}^{-1}(\lfloor \alpha \rfloor \tilde{q}_i + j)}.$$

Importantly, this polynomial basis has the following property $\max_{\tilde{x} \in I_{\tilde{q}}} |Q_{\tilde{a},\tilde{q}}(\tilde{x})| \leq \lfloor \alpha \rfloor^{(d-1)\lfloor \alpha \rfloor}$. We define the estimated polynomial interpolation of g^* for $\tilde{x} \in I_{\tilde{q}}$:

$$\widehat{P}_{\tilde{q}}(\tilde{x}) \doteq \sum_{\substack{\tilde{a} \in \{0,..,M_{l^*}\}^{d-1} \\ \tilde{a}: M_{l^*}^{-1}\tilde{a} \in I_{\tilde{q}}}} T_{l^*,\tilde{a}} Q_{\tilde{q},\tilde{a}}(\tilde{x}).$$

This polynomial interpolation scheme is such that for any $\tilde{a} \in \{0,..,M_{l^*}\}^{d-1}$ with $M_{l^*}^{-1}\tilde{a} \in I_{\tilde{q}}$, we have $\widehat{P}_{\tilde{q}}(M_{l^*}^{-1}\tilde{a}) = T_{l^*,\tilde{a}}$ i.e. we can control exactly the value of the interpolation on the grid. We also define for the entire feature space: $\widehat{P}(\tilde{x}) \doteq \sum_{\tilde{q}} \widehat{P}_{\tilde{q}}(\tilde{x}) \mathbf{1}\{\tilde{x} \in I_{\tilde{q}}\}$. Finally, we define $b_{l^*} = \lambda \lceil \alpha \rceil^{d\lceil \alpha \rceil} M_{l^*}^{-\alpha}$ which is a *bias term* related to the interpolation method we use. Points that are far away enough from the estimate $\widehat{P}$ of the boundary with respect to this bias term are then labeled by the algorithm, as we assign $S^0 \doteq \{x : x_d \leq \widehat{P}(\tilde{x}) - 4b_{l^*}\}$ and $S^1 \doteq \{x : x_d \geq \widehat{P}(\tilde{x}) + 4b_{l^*}\}$ to the labels 0 and 1 respectively. This careful labeling is crucial for the Subroutine to have the desired properties to be used in the aggregation procedure.

The following theorem shows that Algorithm 2 is an acceptable subroutine for the adaptive procedure, as it is *correct* in the sense of Definition 3.

Theorem 4. *Algorithm 2 run on a problem in $\mathcal{P}(\alpha,\kappa)$ with parameters $n, \delta, \lambda, \alpha$ is (δ, Δ_n, n)-correct with Δ_n such that:*

$$\Delta_n \leq 7\lceil \alpha \rceil^{d\lceil \alpha \rceil} 2^\alpha \lambda^{\frac{d-1}{2\alpha(\kappa-1)+d-1}} \left(\frac{\log(n/\delta)}{c_1 n} \right)^{\frac{\alpha}{2\alpha(\kappa-1)+d-1}},$$

where $c_1 = \frac{(\kappa-1)c^2}{400(2\lceil \alpha \rceil)^{d-1}\alpha \log(1/c)\kappa 8^{2(\kappa-1)}}$, and where c is the constant involved in Assumption 2.

The proof of this result can be found in the Appendix in Section 4.

4.2 Learning One-Dimensional Thresholds

In this section, we briefly describe the procedure (derived from recent advances in Yan et al. (2016)) whose objective is to actively find a threshold in the one dimensional problem (see Castro and Nowak (2006); Hanneke (2009); Ramdas and Singh (2013)). This procedure, Algorithm 3 is adaptive

with respect to κ, and is used as a Subroutine for the more involved d dimensional procedure. Fix $\tilde{x} \in [0,1]^{d-1}$, and assume that there exists g^* such that η satisfies Assumptions 1 and 2. In the line-search problem, the goal is to find $x^* = (\tilde{x}, x_d^*)$ such that $g^*(\tilde{x}) = x_d^*$, which is equivalent to finding x_d^* such that $\eta(x^*) \geq 1/2$ and for any $x_d < x_d^*$, $\eta((\tilde{x}, x_d)) < 1/2$. The objective of the Subroutine is to return an interval of length at most ϵ such that the threshold x_d^* is contained in this interval with high-probability, using as few samples as possible.

Algorithm 3 Univariate κ-adaptive procedure (line-search) - Yan et al. (2016)

Input: ϵ, δ
Initialisation: $[L_1, R_1] \leftarrow [0,1]$, $k = 0$, $N = 0$, $K = \lceil \log_2(\frac{1}{2\epsilon}) \rceil$
while $\{k < K - 1\}$ **do**
    $k \leftarrow k+1$, $t_k = 0$, $\delta_k = \frac{\delta}{K2^k}$
    $M_k \leftarrow \frac{L_k + R_k}{2}$, $U_k \leftarrow \frac{R_k - L_k}{4} + L_k$, $V_k \leftarrow \frac{R_k - L_k}{4} + M_k$
    while true do
        $t_k = t_k + 1$; $N = N + 3$
        Request labels in M_k, U_k, V_k, receive $Y_{t_k}(M_k), Y_{t_k}(U_k), Y_{t_k}(V_k)$
        Estimate η for $Z \in \{U_k, M_k, V_k\}$: $\widehat{\eta}_{t_k}(Z) = t_k^{-1} \sum_{i=1}^{t_k} Y_i(Z)$
        if $|\widehat{\eta}_{t_k}(M_k) - 1/2| \geq 2\sqrt{\frac{\log(t_k/\delta_k)}{2t_k}}$ **then**
            if $\widehat{\eta}_{t_k}(M_k) > 1/2$ **then**
                $[L_{k+1}, R_{k+1}] \leftarrow [L_k, M_k]$; **break**
            else
                $[L_{k+1}, R_{k+1}] \leftarrow [M_k, R_k]$; **break**
            end if
        end if
        if $\widehat{\eta}_{t_k}(V_k) - 1/2 \geq 2\sqrt{\frac{\log(t_k/\delta_k)}{2t_k}}$ **and** $1/2 - \widehat{\eta}_{t_k}(U_k) \leq 2\sqrt{\frac{\log(t_k/\delta_k)}{2t_k}}$ **then**
            $[L_{k+1}, R_{k+1}] \leftarrow [U_k, V_k]$; **break**
        end if
    end while
end while
Output: $L_K, R_K, T_K = \frac{R_K - L_K}{2}$ (threshold estimate), $N \leq n$ (budget used)

This procedure is a natural adaptation of the famous bisection method for root-finding of deterministic monotone functions in one-dimension. In the deterministic setting, a simple strategy is to query the middle point of the active segment, and depending on the label returned by the query, continue the procedure with one of the two subintervals - effectively dividing by two the length of the active region with each epoch. In the stochastic setting, the intuition is similar, however, at epoch k, we query successively three active points - the three quartiles of the active segment $[L_k, R_k]$, until we know with a certain confidence δ_k the label of some of these active points. This is done by comparing the empirical mean of the labels observed in each point, with the threshold $1/2$ and a confidence term that depends on the number of times we have queried the active points. By stopping either when M_k or both U_k and V_k can be labeled confidently, this reduces the active segment's size by a factor of 2 at each epoch. The algorithm terminates when it reaches the depth $K = \lceil \log_2(\frac{1}{2\epsilon}) \rceil$ and outputs a final threshold estimate T_K and N the total labeling budget used. The following theorem gives a bound on the number of samples required to return an interval of length at most ϵ such that with high probability the true threshold x_d^* is in this interval.

Theorem 5. *Fix $\tilde{x} \in [0,1]^{d-1}$ and let $x_d^* = g^*(\tilde{x})$ and assume that g^* and η satisfy Assumption 2. Algorithm 3 run with precision ϵ and confidence δ terminates with probability at least $1 - \delta$ and*

returns a threshold estimate T_K such that $|T_K - x_d^*| \leq \epsilon$ using at most N samples with

$$N \leq \begin{cases} 64\big(\log(\frac{1}{\delta}) + \log(\frac{1}{\epsilon})\big)\frac{\log(1/c)}{c^2}\log(\frac{1}{\epsilon}), & \text{if } \kappa = 1 \\[2ex] 200\Big(\log(\frac{1}{\delta}) + \log(\frac{1}{\epsilon})\Big)\frac{\kappa\log(c^{-1})8^{2(\kappa-1)}}{(\kappa-1)c^2}\Big(\big(\frac{1}{\epsilon}\big)^{2(\kappa-1)} - 1\Big), & \text{if } \kappa > 1, \end{cases}$$

where c is the constant involved in Assumption 2.

4.3 Remarks on the Procedures

Optimistic classification. Both Subroutines make optimistic guesses on the labels of the queried points. This is inspired from techniques in the bandit literature (in particular UCB strategies Auer et al. (2002)). In the classification setting, the quantity of interest for a point x is how far this point is from the decision boundary $g^*(x)$, or how far $\eta(x)$ is from $1/2$. By using a confidence term, it is possible to determine with a certain confidence the label of x, or avoid making a potentially wrong guess. In our setting, this observation naturally leads to efficient algorithms that are able to find the decision boundary (up to a certain precision). These optimistic guesses are crucial to show the correctness property required by the aggregation strategy adapted from Locatelli et al. (2017).

Hierarchical zooming. In order to adapt to the noise parameter κ, we keep a hierarchical partitioning of the space which becomes more and more refined. This is related to ideas in the continuous bandit literature, in which the goal is to optimize an unknown function over the domain (see e.g. Kleinberg et al. (2013)). A similar idea was used in Perchet et al. (2013) in the contextual bandit setting and in Locatelli et al. (2017) for active learning, where it is shown that *zooming* strategies naturally adapt to a Tsybakov noise condition.

Improving sample efficiency. We now briefly explain how to modify our procedures to improve the sample efficiency of the adaptive meta-strategy. A key property of our Algorithm 2 is its correctness as demonstrated by Theorem 4. However, our meta-procedure currently only harnesses this property as it aggregates the correct labels output by the different runs of Algorithm 2 which are run independently from one another, leading to potentially wasteful exploration. Instead, Algorithm 2 (run with smoothness parameter α_{i+1}) should only request new labels in points which are within the unlabeled region after the previous aggregation step of the meta-procedure. This can be done efficiently by modifying Subroutine 3 such that it first queries whether the active points M_k, U_k, V_k at round k belong to the aggregated labeled sets s_i^y (for $y \in \{0,1\}$) of the meta-procedure. If some these points have already been confidently labeled, this *correct* label can be re-used directly. Otherwise, the univariate Subroutine requests new labels. Similarly, sample efficiency can be improved within Algorithm 2 itself, by fitting an α_i-smooth boundary at the end of each round indexed by l, *correctly* labeling points on either side of this fitted boundary (with an abstention margin), and aggregating these correct labels (as in the meta-procedure) at each depth. Combined with the previous modification, this ensures that all the information acquired (in the form of correctly labeled sets) is re-used on-the-fly by the non-adaptive procedures. A final modification to improve sample efficiency is to use confidence intervals in Algorithm 3 which exploit the parametric nature of the label distribution $Y(X) \sim \text{Ber}(\eta(X))$ (see e.g. Garivier and Cappé (2011); Kaufmann et al. (2015)). Even though these changes would not improve the rate of convergence of the adaptive meta-procedure in our setting, they should greatly improve its sample efficiency.

Conclusion

We presented in this work the first adaptive strategy for active learning in the boundary fragment setting, resolving a problem that was open since the formulation of this setting in Castro and Nowak (2007), as all known strategies required the knowledge of the characteristic parameters of the problem, which are in general out of reach for practitioners.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: http://dx.doi.org/10. 1023/A:1013689704352. URL `http://dx.doi.org/10.1023/A:1013689704352`.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Rui M Castro. *Active learning and adaptive sampling for non-parametric inference*. PhD thesis, Citeseer, 2007.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. In *44th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2006*, 2006.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. In *International Conference on Computational Learning Theory*, pages 5–19. Springer, 2007.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Information, prediction, and query by committee. In *Advances in neural information processing systems*, pages 483–490, 1993.

Aurélien Garivier and Olivier Cappé. The {KL}-{UCB} algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

S. Hanneke. Adaptive rates of convergence in active learning. *COLT*, 2009.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pages 353–360. ACM, 2007.

Steve Hanneke et al. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 2015.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*, 2013.

V. Koltchinskii. Rademacher complexities and bounding the excess risk of active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

Oleg V Lepski and VG Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546, 1997.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. Adaptivity to noise parameters in nonparametric active learning. *arXiv preprint arXiv:1703.05841*, 2017.

Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13 (Jan):67–90, 2012.

Vianney Perchet, Philippe Rigollet, et al. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

Aaditya Ramdas and Aarti Singh. Algorithmic connections between active learning and stochastic convex optimization. Springer, 2013.

Alexandre Tsybakov. Introduction to nonparametric estimation. 2009.

Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

Liwei Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12(Jul):2269–2292, 2011.

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from imperfect labelers. In *Advances in Neural Information Processing Systems*, pages 2128–2136, 2016.

Appendix A. Proofs

A.1 Proof of Theorem 5

Proof. Fix $\tilde{x} \in [0,1]^{d-1}$. In this proof, with a slight abuse of notation, we write $\eta(Z) \doteq \eta((\tilde{x}, Z))$. Our goal is to find the unique threshold $x_d^* \in [0,1]$ such that we have for any $x_d \geq x_d^*$, $\eta((\tilde{x}, x_d)) \geq 1/2$, and $\eta((\tilde{x}, x_d)) < 1/2$ for any $x_d < x_d^*$ where η is such that Assumption 2 is satisfied for some $\kappa \geq 1$. We will first write the event under which all average estimates used by the algorithm concentrate around their means. For $Z \in [0,1]$ sampled $t \geq 1$ times by the algorithm with $\widehat{\eta}_t(Z) = \sum_{i=1}^{t} Y_t(Z)$ where $Y_t(Z)$ is the t-th observation collected in $(\tilde{x}, Z)$, consider the event:

$$|\widehat{\eta}_t(Z) - \eta(Z)| \leq \sqrt{\frac{\log(1/\delta)}{2t}}.$$

By Chernoff-Hoeffding, this event holds with probability at least $1 - \delta$. We denote G_k the dyadic grid of $[0,1]$ with step size 2^{-k}, i.e. $G_k = \{\frac{i}{2^k}, i \in \{1, ..., 2^k - 1\}\}$. Note that there are $2^k - 1$ points in G_k. Let $K = \lceil \log_2\left(\frac{1}{2\epsilon}\right) \rceil$, $\delta_k = \frac{\delta}{K2^{k+1}}$. We define the event ξ:

$$\xi \doteq \left\{ \forall t, k, i \quad s.t. \quad t \geq 1, k \leq K, Z_{i,k} \in G_k : |\widehat{\eta}_t(Z_{i,k}) - \eta(Z_{i,k})| \leq \sqrt{\frac{\log(\frac{t^2}{\delta_k})}{2t}} \right\},$$

By a union bound, we have:

$$
\begin{aligned}
\mathbb{P}(\bar{\xi}) &\leq \sum_{k \leq K} \sum_{Z_{i,k} \in G_k} \sum_{t \geq 1} \frac{\delta_k}{t^2} \\
&\leq \frac{\pi^2}{6} \sum_{k \leq K} \sum_{Z_{i,k} \in G_k} \frac{\delta}{K2^{k+1}} \\
&\leq \delta,
\end{aligned}
$$

where we use $\sum_{t \geq 1} t^{-2} = \frac{\pi^2}{6} \leq 2$ and the definition of δ_k. This shows that $\mathbb{P}(\xi) \geq 1 - \delta$.

Assume that at the beginning of epoch k, we have $\Delta_k \doteq R_k - L_k = 2^{-k+1}$, and R_k and L_k are such that $x_d^* \in [L_k, R_k]$. As the points U_k, M_k, V_k divide the interval $[L_k, R_k]$ in four subintervals of equal length, and there exists a unique threshold $x_d^* \in [L_k, R_k]$, it implies that there is at most a single point $Z \in \{U_k, M_k, V_k\}$ such that $|Z - x_d^*| < \frac{\Delta_k}{8}$. Consider the case $|U_k - x_d^*| < \frac{\Delta_k}{8}$ - the other cases are handled similarly. We thus have $|M_k - x^*| \geq \frac{\Delta_k}{8}$. This implies by Assumption 2:

$$|\eta(M_k) - \frac{1}{2}| \geq c\left(\frac{\Delta_k}{8}\right)^{\kappa-1}. \tag{3}$$

Without loss of generality, assume that $\widehat{\eta}_{t_k}(M_k) > 1/2$ when the epoch ends for the smallest t_k such that $|\widehat{\eta}_{t_k}(M_k) - 1/2| \geq 2\sqrt{\frac{\log(t_k/\delta_k)}{2t_k}}$. On ξ, we have:

$$\eta(M_k) - \sqrt{\frac{\log(t_k/\delta_k)}{2t_k}} \leq \widehat{\eta}_{t_k}(M_k) \leq \eta(M_k) + \sqrt{\frac{\log(t_k/\delta_k)}{2t_k}} \tag{4}$$

Epoch k ends as soon as $\widehat{\eta}_{t_k}(M_k) - 1/2 \geq 2\sqrt{\frac{\log(t_k/\delta_k)}{2t_k}}$. Combining this condition with Equation (4) brings on ξ:

$$
\begin{aligned}
2\sqrt{\frac{\log(t_k/\delta_k)}{2t_k}} &\leq \widehat{\eta}_{t_k}(M_k) - 1/2 \\
&\leq \eta(M_k) - 1/2 + \sqrt{\frac{\log(t_k/\delta_k)}{2t_k}},
\end{aligned}
$$

15

which implies that $\eta(M_k) \geq \sqrt{\frac{\log(t_k/\delta_k)}{2t_k}} + 1/2 > 1/2$, and we have correctly labeled the point M_k i.e. on ξ, $\mathbb{1}\{\widehat{\eta}_{t_k}(M_k) \geq 1/2\} = \mathbb{1}\{\eta(M_k) \geq 1/2\}$. Equations (4) and (3) together yield that the epoch stops if t_k is such that:

$$3\sqrt{\frac{\log(t_k/\delta_k)}{2t_k}} \leq \eta(M_k) - 1/2, \tag{5}$$

implying the following sufficient condition for epoch k to end: $t_k \geq \frac{9\log(t_k/\delta_k)}{2(\eta(M_k)-1/2)^2}$. Thus, when the epoch ends we have at most:

$$t_k \leq \frac{5\log(t_k/\delta_k)}{(\eta(M_k) - 1/2)^2}.$$

Denote for now $u = (\eta(M_k) - 1/2) \leq 1/2$ as η is bounded in $[0, 1]$ and assume that $t_k \leq \frac{17\log(1/(u^2\delta_k))}{u^2}$. Injecting this in Equation (5) brings that the epoch ends if:

$$t_k \geq \frac{5\log(\frac{17\log(1/(u^2\delta_k))}{u^2\delta_k})}{u^2}. \tag{6}$$

We now check that $\frac{5\log(17\log(1/(u^2\delta_k))/(u^2\delta_k))}{u^2} \leq \frac{17\log(1/(u^2\delta_k))}{u^2}$. This is true if $5\log(\log(1/(u^2\delta_k))) + 5\log(17) \leq 12\log(1/(u^2\delta_k))$. As we have $\delta_k \leq 1$ and $u \leq 1/2$, then $w = 1/(u^2\delta_k) \geq 4$, and one can easily check that $5\log(\log(w)) + 5\log(17) \leq 12\log(w)$ for any $w \geq 4$.

Using Equation (3), we thus have the following upper-bound on t_k:

$$t_k \leq \begin{cases} 17c^{-2}\log\left(\frac{1}{c^2\delta_k}\right), & \text{if } \kappa = 1, \\ 17c^{-2}\log\left(\left(\frac{8}{\Delta_k}\right)^{2(\kappa-1)}\frac{1}{c^2\delta_k}\right)\left(8\Delta_k^{-1}\right)^{2(\kappa-1)}, & \text{if } \kappa > 1. \end{cases}$$

Similarly, we can show that on ξ, we make no mistake in the case $\widehat{\eta}_{t_k}(M_k) < 1/2$ when the epoch stops, and obtain the same bound on t_k. Thus on ξ when epoch k ends, we have identified an interval $[L_{k+1}, R_{k+1}]$ of size $\frac{\Delta_k}{2}$ such that $x^* \in [L_{k+1}, R_{k+1}]$. By recurrence, this shows that on ξ, we have for any $k \leq K$, $x^* \in [L_k, R_k]$ and $\Delta_k = 2^{-k+1}$. We now bound the total budget required for all epochs $k \leq K$ to end on ξ. When the algorithm terminates we have requested N labels with the following upper-bound on N for $\kappa > 1$:

$$\begin{aligned} N &= 3\sum_{k=1}^{K} t_k \\ &\leq 51c^{-2}\sum_{k=1}^{K}\log\left(\left(\frac{8}{\Delta_k}\right)^{2(\kappa-1)}\frac{1}{c^2\delta_k}\right)\left(8\Delta_k^{-1}\right)^{2(\kappa-1)} \\ &\leq 51c^{-2}8^{2(\kappa-1)}\log\left((8\Delta_K^{-1})^{2(\kappa-1)}\frac{K2^{K+1}}{c^2\delta}\right)\sum_{k=1}^{K}2^{2k(\kappa-1)} \tag{7} \\ &\leq 100c^{-2}8^{2(\kappa-1)}\kappa\log\left(\frac{\log_2(1/\epsilon)}{c^2\delta\epsilon}\right)\sum_{k=1}^{K}2^{2k(\kappa-1)} \\ &\leq 100\kappa\log\left(\frac{\log_2(1/\epsilon)}{\epsilon\delta}\right)\log\left(\frac{1}{c}\right)c^{-2}8^{2(\kappa-1)}(\kappa-1)^{-1}(2^{2K(\kappa-1)}-1) \tag{8} \\ &\leq 200\left(\log(\frac{1}{\delta}) + \log(\frac{1}{\epsilon})\right)\kappa\log\left(\frac{1}{c}\right)c^{-2}8^{2(\kappa-1)}(\kappa-1)^{-1}\left((\frac{1}{\epsilon})^{2(\kappa-1)}-1\right) \tag{9} \end{aligned}$$

and for $\kappa = 1$:

$$
\begin{aligned}
N &\leq 16 \log(1/(c^2 \delta_K)) c^{-2} K \\
&\leq 64 \left(\log\left(\frac{1}{\delta}\right) + \log\left(\frac{1}{\epsilon}\right) \right) \log\left(\frac{1}{c}\right) c^{-2} \log\left(\frac{1}{\epsilon}\right).
\end{aligned}
\tag{10}
$$

$\square$

A.2 Proof of Theorem 4

Proof. We first define the event ξ on which all the calls to the Subroutine 3 are successful. Let $\delta_l = \delta(\max(1, \lfloor \alpha \rfloor) 2^{l(d+1)})^{-1}$.

$$
\xi \doteq \left\{ \forall l \geq 1, \forall \tilde{a} \in \{0, ..., M_l\}^{d-1}, |T_{l,\tilde{a}} - g^*(M_l^{-1} \tilde{a})| \leq \epsilon_l \right\},
$$

At depth $l \geq 1$, we launch $(M_l + 1)^{d-1} \leq \max(1, \lfloor \alpha \rfloor) 2^{ld}$ line-search instances with confidence parameter δ_l and precision ϵ_l. Each run, indexed by $\tilde{a} \in \{0, ..., M_l\}^{d-1}$ returns a correct threshold $T_{l,\tilde{a}}$ along the line segment $\mathcal{L}_{\tilde{a}} \doteq \{(M_{l^*}^{-1} \tilde{a}, x_d), x_d \in [0,1]\}$ such that $|T_{l,\tilde{a}} - g^*(M_l^{-1} \tilde{a})| \leq \epsilon_l$ with probability at least $1 - \delta_l$ and using at most $\mathcal{O}\left((\log(1/\epsilon_l) + \log(1/\delta_l)) \epsilon_l^{-2(\kappa - 1)} \right)$ samples (see Theorem 5).

By a union bound, we have $\mathbb{P}(\bar{\xi}) \leq \delta \sum_{l \geq 1} 2^{-l} \leq 2\delta$, which implies that $\mathbb{P}(\xi) \geq 1 - 2\delta$. At depth l, the algorithm performs $(\max(1, \lfloor \alpha \rfloor) 2^l + 1)^{d-1} \leq (2 \lceil \alpha \rceil)^{d-1} 2^{l(d-1)}$ line-searches. By Equation (8) in the proof of Theorem 5, we can upper bound on ξ the total budget that Algorithm 2 uses at depth l, with $\epsilon_l = \lambda 2^{-\alpha l} \geq 2^{-\alpha l}$ as $\lambda \geq 1$:

$$
\begin{aligned}
N_l &\leq (2 \lceil \alpha \rceil)^{d-1} 2^{l(d-1)} \log\left(\frac{2^{l\alpha}}{\delta}\right) 200 \log(1/c) c^{-2} (8/\lambda)^{2(\kappa-1)} \frac{\kappa}{\kappa - 1} 2^{2l\alpha(\kappa-1)} \tag{11} \\
&\leq (2 \lceil \alpha \rceil)^{d-1} 200 \log(1/c) c^{-2} (8/\lambda)^{2(\kappa-1)} \frac{\kappa}{\kappa - 1} \log\left(\frac{2^{l\alpha}}{\delta}\right) 2^{l(2\alpha(\kappa-1)+d-1)} \tag{12}
\end{aligned}
$$

We are now ready to bound the minimal depth l^* reached by the algorithm. We also upper-bound naively l^* by $\log_2(n)$, as the budget is insufficient to query all cells once at this depth for $d \geq 2$. We bound the number of samples required to reach depth l^* on ξ:

$$
\begin{aligned}
\sum_{l=1}^{l^*} N_l &\leq \sum_{l=1}^{l^*} (2 \lceil \alpha \rceil)^{d-1} \log\left(\frac{2^{l\alpha}}{\delta}\right) 200 \log(1/c) c^{-2} (8/\lambda)^{2(\kappa-1)} \frac{\kappa}{\kappa - 1} 2^{l(2\alpha(\kappa-1)+d-1)} \\
&\leq 200 (2 \lceil \alpha \rceil)^{d-1} \log(1/c) c^{-2} (8/\lambda)^{2(\kappa-1)} \frac{\kappa}{\kappa - 1} \log\left(\frac{2^{l^* \alpha}}{\delta}\right) \sum_{l=1}^{l^*} 2^{l(2\alpha(\kappa-1)+d-1)} \\
&\leq 400 (2 \lceil \alpha \rceil)^{d-1} \log(1/c) c^{-2} (8/\lambda)^{2(\kappa-1)} \frac{\kappa \alpha}{\kappa - 1} \log\left(\frac{n}{\delta}\right) 2^{l^*(2\alpha(\kappa-1)+d-1)}. \tag{13}
\end{aligned}
$$

As the algorithm is limited by a maximum budget of n samples, the depth reached on ξ is lower-bounded by the biggest l^* such that:

$$
2^{l^*(2\alpha(\kappa-1)+d-1)} \leq \frac{(\kappa - 1) c^2 \lambda^{2(\kappa-1)}}{400 (2 \lceil \alpha \rceil)^{d-1} \alpha \log(1/c) \kappa 8^{2(\kappa-1)}} \left(\frac{n}{\log(n/\delta)} \right),
$$

which implies that a minimum depth:

$$
l^* \geq \frac{1}{2\alpha(\kappa-1)+d-1} \log_2 \left(\frac{(\kappa - 1) c^2 \lambda^{2(\kappa-1)}}{400 (2 \lceil \alpha \rceil)^{d-1} \alpha \log(1/c) \kappa 8^{2(\kappa-1)}} \left(\frac{n}{\log(n/\delta)} \right) \right) - 1 \tag{14}
$$

17

is reached by the algorithm on ξ. Let $c_1 = \frac{(\kappa-1)c^2}{400(2\lceil\alpha\rceil)^{d-1}\alpha\log(1/c)\kappa 8^{2(\kappa-1)}}$.

Let $\tilde{a} \in \{0, ..., M_{l^*}\}^{d-1}$. On ξ, we have:

$$|T_{l^*,\tilde{a}} - g^*(M_{l^*}^{-1}\tilde{a})| \leq \lambda\left(\frac{M_{l^*}}{\max(1, \lfloor\alpha\rfloor)}\right)^{-\alpha}$$

Note that M_{l^*} is a quantity accessible to the algorithm to construct the confidence bands for the estimation of the boundary, as it is simply the step size of the last completed epoch.

In what follows, we will consider the threshold estimates $(T_{l^*,\tilde{a}})_{\tilde{a}}$ and construct a polynomial approximation of the boundary.

Case 1: $\alpha > 1$. As in Castro and Nowak (2007), we make use of the tensor-product Lagrange polynomials. Let $\tilde{q} \in \{0, ..., \frac{M_{l^*}}{\lfloor\alpha\rfloor} - 1\}^{d-1}$ index the cells:

$$I_{\tilde{q}} \doteq \left[\tilde{q}_1\lfloor\alpha\rfloor M_{l^*}^{-1}, (\tilde{q}_1+1)\lfloor\alpha\rfloor M_{l^*}^{-1}\right] \times ... \times \left[\tilde{q}_{d-1}\lfloor\alpha\rfloor M_{l^*}^{-1}, (\tilde{q}_{d-1}+1)\lfloor\alpha\rfloor M_{l^*}^{-1}\right].$$

These cells partition $[0,1]^{d-1}$ entirely, as we have $M_{l^*} = \lfloor\alpha\rfloor 2^{l^*}$. The tensor-product Lagrange polynomials are defined as follows:

$$Q_{\tilde{q},\tilde{a}}(\tilde{x}) \doteq \prod_{i=1}^{d-1} \prod_{\substack{0 \leq j \leq \lfloor\alpha\rfloor \\ j \neq \tilde{a}_i - \lfloor\alpha\rfloor\tilde{q}_i}} \frac{\tilde{x}_i - M_{l^*}^{-1}(\lfloor\alpha\rfloor\tilde{q}_i + j)}{M_{l^*}^{-1}\tilde{a}_i - M_{l^*}^{-1}(\lfloor\alpha\rfloor\tilde{q}_i + j)}.$$

It is easily shown that (Castro and Nowak (2007); Castro (2007)):

$$\max_{\tilde{x}\in I_{\tilde{q}}}|Q_{\tilde{a},\tilde{q}}(\tilde{x})| \leq \lfloor\alpha\rfloor^{(d-1)\lfloor\alpha\rfloor}. \tag{15}$$

The tensor-product Lagrange polynomial interpolation of g^* for $\tilde{x} \in I_{\tilde{q}}$ is:

$$P_{\tilde{q}}(\tilde{x}) = \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} g^*(M_{l^*}^{-1}\tilde{a})Q_{\tilde{q},\tilde{a}}(\tilde{x}) \tag{16}$$

and we define the polynomial interpolation of g^* for $\tilde{x} \in I_{\tilde{q}}$:

$$\widehat{P}_{\tilde{q}}(\tilde{x}) = \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} T_{l^*,\tilde{a}}Q_{\tilde{q},\tilde{a}}(\tilde{x}). \tag{17}$$

On ξ, since $\epsilon_{l^*} = \left(\frac{M_{l^*}}{\lfloor\alpha\rfloor}\right)^{-\alpha}$:

$$|T_{l^*,\tilde{a}} - g^*(M_{l^*}^{-1}\tilde{a})| \leq \lambda\left(\frac{M_{l^*}}{\lfloor\alpha\rfloor}\right)^{-\alpha}. \tag{18}$$

For any $\tilde{x} \in I_{\tilde{q}}$, the previous equation brings on ξ:

$$
\begin{aligned}
|\widehat{P}_{\tilde{q}}(\tilde{x}) - P_{\tilde{q}}(\tilde{x})| &= \left|\sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} \left(T_{l^*,\tilde{a}} - g^*(M_{l^*}^{-1}\tilde{a})\right)Q_{\tilde{q},\tilde{a}}(\tilde{x})\right| \\
&\leq \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} \lambda\left(\frac{M_{l^*}}{\lfloor\alpha\rfloor}\right)^{-\alpha}|Q_{\tilde{q},\tilde{a}}(\tilde{x})| \\
&\leq \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} \lambda\left(\frac{M_{l^*}}{\lfloor\alpha\rfloor}\right)^{-\alpha}\lfloor\alpha\rfloor^{(d-1)\lfloor\alpha\rfloor} \\
&\leq \lceil\alpha\rceil^{d-1}\lfloor\alpha\rfloor^{(d-1)\lfloor\alpha\rfloor}\lfloor\alpha\rfloor^{\alpha}\lambda M_{l^*}^{-\alpha} \\
&\leq \lceil\alpha\rceil^{d\lceil\alpha\rceil}\lambda M_{l^*}^{-\alpha}, \tag{19}
\end{aligned}
$$

18

where we use Equation (15) in line 4, and upper-bound the number of terms in the sum by $\lceil\alpha\rceil^{d-1}$. We now turn our attention to the approximation properties of $P_{\tilde{q}}$ with respect to g^*, which do not depend on ξ. For any $\tilde{x} \in I_{\tilde{q}}$ and $g^* \in \Sigma(\lambda, \alpha)$, we have:

$$
\begin{aligned}
|P_{\tilde{q}}(\tilde{x}) - g^*(\tilde{x})| &= |P_{\tilde{q}}(\tilde{x}) - \mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(\tilde{x}) + \mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(\tilde{x}) - g^*(\tilde{x})| \\
&\leq |P_{\tilde{q}}(\tilde{x}) - \mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(\tilde{x})| + |\mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(\tilde{x}) - g^*(\tilde{x})| \\
&\leq |P_{\tilde{q}}(\tilde{x}) - \mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(\tilde{x})| + \lambda\Big(\frac{M_{l^*}}{\lfloor\alpha\rfloor}\Big)^{-\alpha},
\end{aligned} \tag{20}
$$

where TP_x is the Taylor polynomial expansion of g in x of degree $\lfloor\alpha\rfloor$. As the Taylor polynomial expansion is of degree $\lfloor\alpha\rfloor$, it is also possible to write $\mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}$ in the tensor-product Lagrange polynomials basis, bringing:

$$
\begin{aligned}
|P_{\tilde{q}}(\tilde{x}) - \mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(\tilde{x})| &= \Big| \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} \Big(g^*(M_{l^*}^{-1}\tilde{a}) - \mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(M_{l^*}^{-1}\tilde{a})\Big) Q_{\tilde{q},\tilde{a}}(\tilde{x})\Big| \\
&\leq \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} |g^*(M_{l^*}^{-1}\tilde{a}) - \mathrm{TP}_{\tilde{q}\lfloor\alpha\rfloor M_{l^*}^{-1}}(M_{l^*}^{-1}\tilde{a})| |Q_{\tilde{q},\tilde{a}}(\tilde{x})|| \\
&\leq \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} \lambda\Big(\frac{M_{l^*}}{\lfloor\alpha\rfloor}\Big)^{-\alpha} |Q_{\tilde{q},\tilde{a}}(\tilde{x})| \\
&\leq \sum_{\tilde{a}:M_{l^*}^{-1}\tilde{a}\in I_{\tilde{q}}} \lambda\Big(\frac{M_{l^*}}{\lfloor\alpha\rfloor}\Big)^{-\alpha} \lfloor\alpha\rfloor^{(d-1)\lfloor\alpha\rfloor} \\
&\leq \lceil\alpha\rceil^{d-1} \lfloor\alpha\rfloor^{(d-1)\lfloor\alpha\rfloor} \lfloor\alpha\rfloor^{\alpha} \lambda M_{l^*}^{-\alpha} \\
&\leq \lceil\alpha\rceil^{d\lceil\alpha\rceil} \lambda M_{l^*}^{-\alpha},
\end{aligned}
$$

where the third line is obtained by using Assumption 1 as g^* is α-smooth. Combining this with Equation (20) yields the following inequality:

$$
|P_{\tilde{q}}(\tilde{x}) - g^*(\tilde{x})| \leq 2\lceil\alpha\rceil^{d\lceil\alpha\rceil} \lambda M_{l^*}^{-\alpha}. \tag{21}
$$

We are now ready to conclude the proof. Combining Equations (19) and (21) allows us to write:

$$
\begin{aligned}
|\widehat{P}_{\tilde{q}}(\tilde{x}) - g^*(\tilde{x})| &\leq |\widehat{P}_{\tilde{q}}(\tilde{x}) - P_{\tilde{q}}(\tilde{x})| + |P_{\tilde{q}}(\tilde{x}) - g^*(\tilde{x})| \\
&\leq 3\lceil\alpha\rceil^{d\lceil\alpha\rceil} \lambda M_{l^*}^{-\alpha},
\end{aligned}
$$

which brings immediately with $b_{l^*} = \lceil\alpha\rceil^{d\lceil\alpha\rceil} \lambda M_{l^*}^{-\alpha}$ as defined in the algorithm:

$$
0 < b_{l^*} \leq (\widehat{P}_{\tilde{q}}(\tilde{x}) + 4b_{l^*}) - g^*(\tilde{x}) \leq 7b_{l^*}.
$$

This implies directly the following inclusions on ξ:

$$
\{x : x_d \geq g^*(\tilde{x}) + 7b_{l^*}\} \subset S^1 \subset \{x : x_d > g^*(\tilde{x})\}
$$

Through similar considerations, it is easily shown that on ξ, we also have:

$$
\{x : x_d \leq g^*(\tilde{x}) - 7b_{l^*}\} \subset S^0 \subset \{x : x_d < g^*(\tilde{x})\}.
$$

This shows that the procedure is $(n, \delta, \Delta_{l^*})$-correct with:

$$
\Delta_{l^*} \leq 7\lceil\alpha\rceil^{d\lceil\alpha\rceil} 2^\alpha \lambda^{\frac{d-1}{2\alpha(\kappa-1)+d-1}} \Big(\frac{\log(n/\delta)}{c_1 n}\Big)^{\frac{\alpha}{2\alpha(\kappa-1)+d-1}}.
$$

Case 2: $\alpha \leq 1$. We simply use a constant approximation directly on the cells:

$$C_{\tilde{h}} \doteq \left[\tilde{h}_1 M_{l^*}^{-1}, (\tilde{h}_1 + 1)M_{l^*}^{-1}\right] \times ... \times \left[\tilde{h}_{d-1} M_{l^*}^{-1}, (\tilde{h}_{d-1} + 1)M_{l^*}^{-1}\right],$$

indexed by $\tilde{h} \in \{0, ..., M_{l^*} - 1\}$. For $\alpha \leq 1$, the assumption on the smoothness of the boundary simply yields for any $\tilde{h} \in \{0, ..., M_{l^*} - 1\}$ and any $\tilde{x}, \tilde{y} \in C_{\tilde{h}}$:

$$|g^*(\tilde{x}) - g^*(\tilde{y})| \leq \lambda \|\tilde{x} - \tilde{y}\|_\infty^\alpha \leq \lambda M_{l^*}^{-\alpha}. \tag{22}$$

Note that for $\alpha \leq 1$, we have $b_{l^*} = \lambda M_{l^*}^{-\alpha}$, as we have $\lceil \alpha \rceil = 1$. Equation (18) and Equation (22) yield for any $\tilde{x} \in C_{\tilde{h}}$:

$$0 < b_{l^*} \leq T_{l^*, \tilde{h}} + 2b_{l^*} - g^*(\tilde{x}) \leq 4b_{l^*},$$

which shows the $(n, \delta, \Delta_{l^*})$ correctness of the procedure with:

$$\Delta_{l^*} \leq 2^\alpha 5\lambda^{\frac{d-1}{2\alpha(\kappa-1)+d-1}} \left(\frac{\log(n/\delta)}{c_1 n}\right)^{\frac{\alpha}{2\alpha(\kappa-1)+d-1}}$$

$\square$

A.3 Proof of Proposition 1 and Theorem 1

Proof. The proof follows from arguments in Locatelli et al. (2017), adapted to this different notion of correctness.

Set as in Algorithm 1:

$$n_0 = \frac{n}{\lfloor \log(n) \rfloor^2}, \quad \delta_0 = \frac{\delta}{\lfloor \log(n) \rfloor^2}, \quad \text{and} \quad \alpha_i = \frac{i}{\lfloor \log(n) \rfloor}.$$

In Algorithm 1, the Subroutine is launched $\lfloor \log(n) \rfloor^2$ times on $\lfloor \log(n) \rfloor^2$ independent subsamples of size n_0. We index each launch by i, which corresponds to the launch with smoothness parameter α_i. Let i^* be the largest integer $1 \leq i \leq \lfloor \log(n) \rfloor^2$ such that $\alpha_i \leq \alpha$.

Since the Subroutine is strongly $(\delta_0, \Delta_\alpha, n_0)$-correct for any $\alpha \in [\lfloor \log(n) \rfloor^{-1}, \lfloor \log(n) \rfloor]$, it holds by Definition 3 that for any $i \leq i^*$, with probability larger than $1 - \delta_0$

$$\left\{x \in [0,1]^d : x_d - g^*(\tilde{x}) > \Delta_{\alpha_i}\right\} \subset S_i^1 \subset \left\{x \in [0,1]^d : x_d - g^*(\tilde{x}) > 0\right\}$$

and

$$\left\{x \in [0,1]^d : g^*(\tilde{x}) - x_d > \Delta_{\alpha_i}\right\} \subset S_i^0 \subset \left\{x \in [0,1]^d : g^*(\tilde{x}) - x_d > 0\right\}.$$

So by an union bound we know that with probability larger than $1 - \lfloor \log(n) \rfloor^2 \delta_0 = 1 - \delta$, the above equations hold jointly for any $i \leq i^*$.

This implies that with probability larger than $1 - \delta$, we have for any $i' \leq i \leq i^*$, and for any $y \in \{0,1\}$, that

$$S_i^y \cap s_{i'}^{1-y} = \emptyset,$$

i.e. the labeled regions of $[0,1]^d$ are not in disagreement for any two runs of the algorithm that are indexed with parameters smaller than i^*. So we know that just after iteration i^* of Algorithm 1, we have with probability larger than $1 - \delta$, that for any $y \in \{0,1\}$

$$\bigcup_{i \leq i^*} S_i^y \subset s_{i^*}^y.$$

Since the sets s_i^y are strictly growing but disjoint with the iterations i by definition of Algorithm 1 (i.e. $s_i^k \subset s_{i+1}^k$ and $s_i^k \cap s_i^{1-k} = \emptyset$), it holds in particular that with probability larger than $1 - \delta$ and for any $y \in \{0,1\}$

$$\bigcup_{i \leq i^*} S_i^y \subset s_{\lfloor \log(n)\rfloor^2}^y \quad \text{and} \quad s_{\lfloor \log(n)\rfloor^2}^y \cap s_{\lfloor \log(n)\rfloor^2}^{1-y} = \emptyset.$$

This finishes the proof of Proposition 1.

By Proposition 1, Algorithm 1 is weakly-$(\delta_0, \Delta_{\alpha_i}, n_0)$ correct for the largest i such that $\alpha_i \leq \alpha$, with Δ_{α_i} bounded as:

$$\Delta_{\alpha_i} \leq 7\lceil \alpha_i\rceil^{d\lceil \alpha_i\rceil} 2^{\alpha_i} \lambda \left(\frac{\log^3(n/\delta)}{c_1 n}\right)^{\frac{\alpha_i}{2\alpha_i(\kappa-1)+d-1}},$$

with $c_1 = \frac{(\kappa-1)c^2}{400(2\lceil\alpha\rceil)^{d-1}\alpha\log(1/c)\kappa 8^{2(\kappa-1)}}$.

By definition of α_i, which is on a grid of step $\lfloor \log(n)\rfloor^{-1}$, we have:

$$\alpha - \frac{1}{\lfloor \log(n)\rfloor} \leq \alpha_i \leq \alpha.$$

This yields for the exponent in the rate:

$$-\frac{\alpha_i}{2\alpha_i(\kappa-1)+d-1} \leq -\frac{\alpha}{2\alpha(\kappa-1)+d-1} + \frac{\lfloor \log(n)\rfloor^{-1}}{2\alpha(\kappa-1)+d-1}.$$

The result follows by noticing that:

$$n^{\frac{1}{\lfloor \log(n)\rfloor(2\alpha(\kappa-1)+d-1)}} \leq \exp\left(\frac{\log(n)}{\lfloor \log(n)\rfloor(d-1)}\right)$$

and thus this term only affects the rate as a multiplicative constant that does not depend on n, δ and λ. $\qquad\square$

A.4 Proof of Theorem 2

Proof. The basic argument is based on standard applications of Fano's inequality, in particular on a useful form given in Theorem 2.5 in Tsybakov (2009) (which we recall hereunder). The main work is in constructing a suitable family of problems satisfying the conditions of Theorem 6 and matching our distributional requirements.

Theorem 6 (Tsybakov). *Let $\mathcal{H}$ be a class of models, $d : \mathcal{H} \times \mathcal{H} \to \mathbb{R}^+$ a pseudo-metric, and $\{P_\eta, \eta \in \mathcal{H}\}$ a collection of probability measures associated with $\mathcal{H}$. Assume there exists a subset $\{\eta_0, ..., \eta_M\}$ of $\mathcal{H}$ such that:*

1. *$d(\eta_i, \eta_j) \geq 2s > 0$ for all $0 \leq i < j \leq M$*

2. *P_{η_i} is absolutely continuous with respect to P_{η_0} for every $0 < i \leq M$*

3. *$\frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(P_{\eta_i}, P_{\eta_0}) \leq \alpha \log(M)$, for $0 < \alpha < \frac{1}{8}$*

then

$$\inf_{\hat\eta} \sup_{\eta\in\mathcal{H}} P_\eta\big(d(\hat\eta, \eta) \geq s\big) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log(M)}}\right),$$

where the infimum is taken over all possible estimators of η based on a sample from P_η.

Let $\alpha > 0$ and $d \in \mathbb{N}$, $d > 1$. For $x \in \mathbb{R}^d$, we write $x = (x^{(1)}, \cdots, x^{(d)})$ and $x^{(i)}$ denotes the value of the i-th coordinate of x. As previously, for $x \in [0,1]^d$, we use the notation $\tilde{x} = (x^{(1)}, \ldots, x^{(d-1)})$.

Consider the grid of $[0,1]^{d-1}$ of step size $2\Delta^{1/\alpha}$, $\Delta > 0$. There are

$$K = 2^{1-d} \Delta^{(1-d)/\alpha},$$

disjoint hypercubes in this grid, and we write them $(H'_k)_{k \leq K}$. For $k \leq K$, let $\tilde{x}_k$ be the barycenter of H'_k.

We now define the partition of $[0,1]^d$:

$$[0,1]^d = \bigcup_{k=1}^{K} H_k = \bigcup_{k=1}^{K} (H'_k \times [0,1]),$$

where $H_k = (H'_k \times [0,1])$ is an hyper-rectangle corresponding to H'_k - these are hyper-rectangles of side $2\Delta^{1/\alpha}$ along the first $(d-1)$ dimensions, and side 1 along the last dimension.

We define f for any $z \in [\frac{1}{2}\Delta^{1/\alpha}, \Delta^{1/\alpha}]$ as

$$f(z) = \begin{cases} C_{\lambda,\alpha} 4^{\alpha-1} \left(\Delta^{1/\alpha} - z \right)^{\alpha}, & \text{if } \frac{3}{4}\Delta^{1/\alpha} < z \leq \Delta^{1/\alpha} \\ C_{\lambda,\alpha} \left(\frac{\Delta}{2} - 4^{\alpha-1} \left(z - \frac{1}{2}\Delta^{1/\alpha} \right)^{\alpha} \right), & \text{if } \frac{1}{2}\Delta^{1/\alpha} \leq z \leq \frac{3}{4}\Delta^{1/\alpha}, \end{cases}$$

where $C_{\lambda,\alpha} > 0$ is a small constant that depends only on α, λ.

For $k \leq K$, and for any $\tilde{x} \in H'_k$, we write

$$\Psi_k(\tilde{x}) \begin{cases} \frac{C_{\lambda,\alpha}\Delta}{2}, & \text{if } |\tilde{x} - \tilde{x}_k|_2 \leq \frac{\Delta^{1/\alpha}}{2} \\ 0, & \text{if } |\tilde{x} - \tilde{x}_k|_2 \geq \Delta^{1/\alpha} \\ f(|\tilde{x} - \tilde{x}_k|), & \text{otherwise}, \end{cases}$$

which we use to define g_{k_s} over the same domain, for $s \in \{-1,1\}$:

$$g_{k,s}(\tilde{x}) = \frac{1}{2} + s\Psi_k(\tilde{x})$$

f is such that $f(\frac{1}{2}\Delta^{1/\alpha}) = \frac{C_{\lambda,\alpha}\Delta}{2}$, and $f(\Delta^{1/\alpha}) = 0$. Moreover, it is (λ, α)-Hölder on $[\frac{1}{2}\Delta^{1/\alpha}, \Delta^{1/\alpha}]$ for $C_{\lambda,\alpha}$ small enough (depending only on α, λ), and such that all its derivatives are 0 in $\frac{1}{2}\Delta^{1/\alpha}$, $\Delta^{1/\alpha}$. By definition of $\Psi_{k,s}$, it holds that $g_{k,s}$ is in $\Sigma(\lambda, \alpha)$ restricted to H'_k.

We now define $\eta_{k,s}$ for $x \in H_k$:

$$\eta_{k,s}(x) = \begin{cases} c|x_d - g_{k,s}(\tilde{x}) + 2\Psi_k(\tilde{x})|^{\kappa-1} & \text{if } s(x_d - g_{k,s}(\tilde{x})) > 2\Psi_k(\tilde{x}) \\ c|x_d - g_{k,s}(\tilde{x})|^{\kappa-1} & \text{otherwise}. \end{cases}$$

We see immediately by definition of $\eta_{k,s}$ that it satisfies Assumption 2, and that $\eta_{k,-1}(x) = \eta_{k,1}(x)$ for $\{x : |x_d - 1/2| \geq \Psi_k(\tilde{x})\}$ (i.e. $\eta_{k,s}$ only depends on s in a small band around the decision boundary).

For $\sigma \in \{-1,1\}^K$, we define for any $\tilde{x} \in [0,1]^{d-1}$ the function

$$g_{\sigma}^*(\tilde{x}) = \sum_{k \leq K} g_{k,\sigma_k}(\tilde{x}) \mathbf{1}\{\tilde{x} \in H'_k\}.$$

Note that since each $g_{k,s}$ is in $\Sigma(\lambda, \alpha)$ restricted to H'_k, and by definition of $g_{k,s}$ at the borders of each H'_k, it holds that g_{σ}^* is in $\Sigma(\lambda, \alpha)$ on $[0,1]^{d-1}$.

We now define the marginal distribution $\mathbb{P}_X$ of X. To simplify notations, we first define for any $x \in H_k$: $D_k(x) = \min(|x_d - g_{k,1}(\tilde{x})|, |x_d - g_{k,-1}(\tilde{x})|)$ and $D(x) = \sum_{k=1}^{K} D_k(x)\mathbf{1}\{x \in H_k\}$. This is simply the distance from x to the closest possible location of the boundary, and it does not depend on $s \in \{-1, 1\}$. We define p_k for $x \in H_k$ for $\kappa' > \kappa - 1$:

$$p_k(x) = \begin{cases} C_1 D_k(x)^{\kappa' - \kappa} & \text{if } D_k(x) \le \Delta_0 \\ C_2 & \text{otherwise.} \end{cases}$$

This allows us to define the density:

$$p(x) = \sum_{k=1}^{K} p_k(x)\mathbf{1}\{x \in H_k\},$$

where the constants C_1 and C_2 are chosen such that Assumption 3 is satisfied and p integrates to 1 over $[0, 1]^d$.

Finally, for any $\sigma \in \{-1, +1\}^K$, we define P_{η_σ} as the measure of the data in our setting when the density of $\mathbb{P}_X$ is p, and where the regression function $\mathbb{P}_{Y|X}$ providing the distribution of the labels is η_σ. By a slight abuse of notation, we write $P_\sigma = P_{\eta_\sigma}$. We write

$$\mathcal{H} = \{\eta_\sigma : \sigma \in \{-1, +1\}^K\}.$$

For any element η_σ of $\mathcal{H}$, P_σ satisfies Assumptions, 1, 2 and 3 by construction.

We define $P_{\sigma,n}$ the joint distribution $(X_i, Y_i)_{i=1}^n$ of samples collected by any (possibly active) fixed sampling strategy Π_n under P_σ, where $\Pi_n = \{\pi_i\}_{i \le n}$, and $\pi_t(x, \{(X_i, Y_i)\}_{i<t})$ is the sampling strategy at time t that depends on the samples collected up to time t. π_t defines the sampling rule $\pi_t(x, \{(X_i, Y_i)\}_{i<t}) = P_{\pi,\sigma}(X_t = x|(X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}))$, for any $x \in [0, 1]^d$. We remark here that this sampling mechanism may depend on $\mathbb{P}_X$, which is why we have constructed $\mathbb{P}_X$ such that it does not depend on σ. This is crucial for Proposition 3 (from Castro and Nowak (2007)) to hold. As $\mathbb{P}_X$ does not depend on σ, we have immediately that $\forall i \le M$, $P_{\sigma_i,n}$ is absolutely continuous with respect to $P_{\sigma_0,n}$.

Proposition 2 (Gilbert-Varshamov). *For $K \ge 8$ there exists a subset $\{\sigma_0, ..., \sigma_M\} \subset \{-1, 1\}^K$ such that $\sigma_0 = \{1, ..., 1\}$, $\rho(\sigma_i, \sigma_j) \ge \frac{K}{8}$ for any $0 \le i < j \le M$ and $M \ge 2^{K/8}$, where ρ stands for the Hamming distance between two sets of length K.*

We denote $\mathcal{H}' \doteq \{\eta_{\sigma_0}, \cdots, \eta_{\sigma_M}\}$ a subset of $\mathcal{H}$ of cardinality $M \ge 2^{K/8}$ with $K \ge 8$ such that for any $1 \le k < j \le M$, we have $\rho(\sigma_k, \sigma_j) \ge K/8$. We know such a subset exists by Proposition 2.

Proposition 3 (Castro and Nowak). *For any $\sigma \in \mathcal{H}$ such that $\sigma \ne \sigma_0$ and Δ small enough such that $\eta_\sigma, \eta_{\sigma_0}$ take values only in $[1/5, 4/5]$ and $\mathbb{P}_X$ does not depend on σ, we have:*

$$\mathrm{KL}(P_{\sigma,n}||P_{\sigma_0,n}) \le 7n \max_{x \in [0,1]^d} (\eta_\sigma(x) - \eta_{\sigma_0}(x))^2.$$

where $\mathrm{KL}(.||.)$ is the Kullback-Leibler divergence between two-distributions, and $P_{\sigma,n}$ stands for the joint distribution $(X_i, Y_i)_{i=1}^n$ of samples collected by any (possibly active) fixed sampling strategy under P_σ.

This proposition is a consequence of the analysis in Castro and Nowak (2008) (Theorem 1 and 3, and Lemma 1). A proof can be found in Minsker (2012).

By Definition of the η_σ, we know that $\max_{x \in [0,1]^d} |\eta_\sigma(x) - \eta_{\sigma_0}(x)| \leq c(2C_{\lambda,\alpha}\Delta)^{\kappa-1}$, and so Proposition 3 implies that for any $\sigma \in \mathcal{H}'$:

$$
\begin{aligned}
\mathrm{KL}(P_{\sigma,n}||P_{\sigma_0,n}) &\leq 7n \max_{x \in [0,1]^d} (\eta_\sigma(x) - \eta_{\sigma_0}(x))^2 \\
&\leq 7nc^2(2C_{\lambda,\alpha})^{2(\kappa-1)}\Delta^{2(\kappa-1)}.
\end{aligned}
$$

So we have :

$$
\frac{1}{M} \sum_{\sigma \in \mathcal{H}'} \mathrm{KL}(P_{\sigma,n}||P_{\sigma_0,n}) \leq 7nc^2(2C_{\lambda,\alpha})^{2(\kappa-1)}\Delta^{2(\kappa-1)} < \frac{K}{8^2} \leq \frac{\log(|\mathcal{H}'|)}{8},
$$

for n larger than a constant that depends only on α, λ, and setting

$$
\Delta = C_3 n^{-\alpha/(2(\kappa-1)\alpha+d-1)},
$$

as $K = C_4 \Delta^{(1-d)/\alpha}$. This implies that for this choice of Δ, the third condition in Theorem 6 is satisfied.

Finally, we define the pseudo-metric as follows:

$$
d(\eta, \eta') = \int_{x \in [0,1]^d} \mathbf{1}\{\mathrm{sign}(\eta(x) - 1/2) \neq \mathrm{sign}(\eta'(x) - 1/2)\} D(x)^{\kappa-1} p(x) \mathrm{d}x.
$$

For $\sigma, \sigma' \in \mathcal{H}'$, we have:

$$
\begin{aligned}
d(\eta_\sigma, \eta_{\sigma'}) &= \int_{x \in [0,1]^d} \mathbf{1}\{\mathrm{sign}(\eta_\sigma(x) - 1/2) \neq \mathrm{sign}(\eta_{\sigma'}(x) - 1/2)\} D(x)^{\kappa'-1} \mathrm{d}x. \\
&= C_5 \rho(\sigma, \sigma') \int_{\tilde{x} \in H_1'} \left(\int_{|x_d - 1/2| \leq \Psi_1(\tilde{x})} \min(|x_d - g_{1,1}(\tilde{x})|, |x_d - g_{1,-1}(\tilde{x})|)^{\kappa'-1} \mathrm{d}x_d \right) \mathrm{d}\tilde{x} \\
&= 2C_5 \rho(\sigma, \sigma') \int_{\tilde{x} \in H_1'} \left(\int_{1/2}^{1/2+\Psi_k(\tilde{x})} |x_d - g_{1,1}(\tilde{x})|^{\kappa'-1} \mathrm{d}x_d \right) \mathrm{d}\tilde{x} \\
&\geq 2C_5 \rho(\sigma, \sigma') \int_{|\tilde{x}-x_k|_2 \leq \frac{\Delta^{1/\alpha}}{2}} \left(\int_{1/2}^{1/2+\frac{C_{\lambda,\alpha}\Delta}{2}} \left| x_d - \frac{C_{\lambda,\alpha}\Delta}{2} + \frac{1}{2} \right|^{\kappa'-1} \mathrm{d}x_d \right) \mathrm{d}\tilde{x} \\
&\geq C_6 \rho(\sigma, \sigma') \Delta^{(d-1)/\alpha} \Delta^{\kappa'} \\
&\geq C_7 \Delta^{\kappa'},
\end{aligned}
$$

where we use the definition of p in the first line, the definition of η_σ and $\rho(\sigma, \sigma')$ and Fubini's theorem in the second line, and the lower bound on $\rho(\sigma, \sigma')$ by definition of $\mathcal{H}'$ in the last line.

All assumptions in Theorem 6 are thus satisfied with $s = C_7 \Delta^{\kappa'}$ and $\Delta = C_3 n^{-\alpha/(2(\kappa-1)\alpha+d-1)}$. For any $\eta_\sigma \in \mathcal{H}'$, and any $\hat{\eta} : [0,1]^d \to [0,1]$:

$$
\begin{aligned}
d(\hat{\eta}_n, \eta_\sigma) &= \int_{x \in [0,1]^d} \mathbf{1}\{\mathrm{sign}(\hat{\eta}_n(x) - 1/2) \neq \mathrm{sign}(\eta_\sigma(x) - 1/2)\} D(x)^{\kappa-1} p(x) dx \\
&\leq c^{-1} \int_{x \in [0,1]^d} \mathbf{1}\{\mathrm{sign}(\hat{\eta}_n(x) - 1/2) \neq \mathrm{sign}(\eta_\sigma(x) - 1/2)\} |1 - 2\eta_\sigma(x)| p(x) dx \\
&\leq \frac{R_{P_\sigma}(\hat{\eta}_n) - R_{P_\sigma}(\eta_\sigma)}{c}
\end{aligned}
$$

where we use in the second line the fact that η_σ satisfies Assumption 2 with constant c, and thus under P_σ, we have $d(\hat{\eta}, \eta_\sigma) \leq c^{-1}\mathcal{E}_{P_\sigma}(\hat{\eta})$. We can now apply Theorem 6, which yields for any fixed sampling strategy π_n as defined previously:

$$\inf_{\hat{\eta}_n} \sup_{\eta_\sigma \in \mathcal{H}} P_{\sigma,n}\left(\mathcal{E}(\hat{\eta}_n) \geq C_8 n^{-\kappa'\alpha/(2\alpha(\kappa-1)+d-1)}\right) \geq C_9,$$

where C_9 is a small universal constant. We conclude by applying Markov's inequality, and taking the infimum over (possibly active) sampling strategies Π_n (as this holds for any strategy Π_n). $\qquad\square$

A.5 Proof of Theorem 3 (passive lower bound)

Proof. In the passive setting, the proof is the same but we need a different bound on the quantity:

$$\text{KL}(P_{\sigma,n}||P_{\sigma_0,n}) = n \int_{x:\eta_\sigma(x)\neq\eta_{\sigma_0}(x)} d_{\text{KL}}(\eta_\sigma(x), \eta_{\sigma_0}(x))p(x)\mathrm{d}x,$$

where $d_{\text{KL}}(p,q)$ stands for the Kullback-Leibler divergence between two Bernoulli distributions of parameters p, q. Instead, we bound it as:

$$\text{KL}(P_{\sigma,n}||P_{\sigma_0,n}) \leq nC_{10}\Delta^{\kappa'+\kappa-1},$$

using $d_{\text{KL}}(\eta_\sigma(x), \eta_{\sigma_0}(x)) \leq C_{11}\Delta^{2(\kappa-1)}$ by Pinsker's inequality for $\eta(x) \in [1/5, 4/5]$, and the definition of $p(x)$. We conclude by setting $\Delta = C_{12}n^{-\alpha/(\alpha(\kappa+\kappa'-1)+d-1)}$ to satisfy the assumptions of Theorem 6. $\qquad\square$