Proceedings of Machine Learning Research vol 65:1–34, 2017

Adaptivity to Noise Parameters in Nonparametric Active Learning

Andrea Locatelli LOCATELL@UNI-POTSDAM.DE
Department of Mathematics, University of Potsdam, Germany

Alexandra Carpentier CARPENTIER@UNI-POTSDAM.DE
Department of Mathematics, University of Potsdam, Germany

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

Abstract

This work addresses various open questions in the theory of active learning for nonparametric classification. Our contributions are both statistical and algorithmic:

- We establish new minimax-rates for active learning under common *noise conditions*. These rates display interesting transitions – due to the interaction between noise *smoothness and margin* – not present in the passive setting. Some such transitions were previously conjectured, but remained unconfirmed.
- We present a generic algorithmic strategy for adaptivity to unknown noise smoothness and margin; our strategy achieves optimal rates in many general situations; furthermore, unlike in previous work, we avoid the need for *adaptive confidence sets*, resulting in strictly milder distributional requirements.

Keywords: Active learning, Nonparametric classification, Noise conditions, Adaptivity.

1. Introduction

The nonparametric setting in classification allows for a generality which has so far provided remarkable insights on how the interaction between distributional parameters controls learning rates. In particular the interaction between feature $X \in \mathbb{R}^d$ and label $Y \in \{0, 1\}$ can be parametrized into *label-noise* regimes that clearly interpolate between hard and easy problems. This theory is now well developed for *passive learning*, i.e., under i.i.d. sampling, however for *active learning* – where the learner actively chooses informative samples – the theory is still evolving. Our goals in this work are both statistical and algorithmic, the common thrust being to better understand how label-noise regimes control the active setting and induce performance gains over the passive setting.

An initial nonparametric result of Castro and Nowak (2008) considers situations where the Bayes decision boundary $\{x : \mathbb{E}[Y|X = x] = 1/2\}$ is given by a *smooth* curve which bisects the X space. The work yields nontrivial early insights into nonparametric active learning by formalizing a situation where active rates are significantly faster than their passive counterpart.

More recently, Minsker (2012a) considered a different nonparametric setting, also of interest here. Namely, rather than assuming a smooth boundary between the classes, the joint distribution of the data $\mathbb{P}_{X,Y}$ is characterized in terms of the *smoothness* α of the regression function $\eta(x) \doteq \mathbb{E}[Y|X = x]$; this setting has the appeal of allowing more general decision boundaries.

Furthermore, following Audibert and Tsybakov (2007), the *noise level* in Y, i.e., the likelihood that $\eta(X)$ is close to $1/2$, is captured by a *margin* parameter β. Restricting attention to the case $\alpha \leq 1$ (Hölder continuity) and $\alpha\beta \leq d$, Minsker (2012a) shows striking improvements in the active rates over passive rates, including an interesting phenomenon for the active rate at the perimeter $\alpha\beta = d$. More precisely, under certain technical conditions, the minimax rate (excess error over the Bayes classifier) is of the form $n^{-\alpha(\beta+1)/(2\alpha+d-\alpha\beta)}$, where n is the number of samples requested. In contrast, the passive rate is $n^{-\alpha(\beta+1)/(2\alpha+d)}$, i.e., the dependence on dimension d is greatly reduced with large $\alpha\beta$, down to (nearly) *no dependence*[1] on d when $\alpha\beta = d$. For the case $\alpha > 1$, quite interestingly, later work Minsker (2012b) obtains a different upper-bound of the form $n^{-\alpha(\beta+1)/(2\alpha+d-\beta)}$, i.e., the dependence on d is now only reduced by the noise term β rather than by $\alpha\beta$ as when $\alpha \leq 1$. While there was no matching lower-bound, both Minsker (2012a,b) conjecture that this rate is tight, i.e., that there might indeed be a phase transition at $\alpha \geq 1$. Nevertheless, the evolving picture is one where the interaction between α, β and d seems essential in active learning.

Thus, many natural questions remain open in the present setting (of Audibert and Tsybakov (2007) and Minsker (2012a)). First, statistical rates remain unclear in various regimes: when Hölder smoothness $\alpha > 1$, when $\alpha\beta > d$, or when the marginal distribution $\mathbb{P}_X$ is far from being uniform on $[0, 1]^d$ (this is required in Minsker (2012a) and in the earlier setting of Castro and Nowak (2008)). Furthermore, a nontrivial algorithmic problem remains: a natural active strategy is to query Y at x only when we lack confidence in the label estimate at x, i.e., when $\eta(x) \doteq \mathbb{E}[Y|x]$ is deemed close to $1/2$; this seemingly requires tight assessments of the *confidence* in estimates of $\eta(x)$, however, such confidence assessment is challenging without a priori knowledge of distributional parameters such as the smoothness α of η. In fact, this is a challenge in any nonparametric setting, and Castro and Nowak (2008) for instance simply assume knowledge of relevant parameters. In our particular setting, the only known procedures of Minsker (2012a,b) have to resort to restrictive conditions [2] outside of which *adaptive and honest*[3] confidence sets do not exist (see negative results of Robins et al. (2006); Cai et al. (2006); Genovese and Wasserman (2008); Hoffmann and Nickl (2011); Bull and Nickl (2013); Carpentier (2013)). We present a simple strategy that bypasses adaptive and honest confidence sets, and therefore avoids ensuing restrictive conditions.

Statistical results. The present work expands on the existing theory of nonparametric active learning in many directions, and confirms new interesting transitions in achievable rates induced by regimes of interaction between distributional parameters α, β, d and the marginal $\mathbb{P}_X$. We outline some noteworthy such transitions below. We assume as in prior work that supp $(\mathbb{P}_X) \subset [0, 1]^d$:

- For $\alpha > 1$, $\mathbb{P}_X$ nearly uniform, Minsker (2012a) conjectured that the minimax rates for active learning changes to $n^{-\alpha(\beta+1)/(2\alpha+[d-\beta]_+)}$, i.e., $[d-\beta]_+$ should appear in the denominator rather than by $[d-\alpha\beta]_+$. We show that this rate is indeed tight in relevant cases: the *upper-bound* $n^{-\alpha(\beta+1)/(2\alpha+[d-\beta]_+)}$ is attained by our algorithm for any $\alpha \geq 1, \beta \geq 0$, while we establish a matching lower-bound when $\beta = 1$; in other words, a better upper-bound is impossible without additional assumptions on β. This however leaves open the possibility of a much richer set of transitions characterized by β. We note that no such transition at $\alpha > 1$ is known in the passive

[1] In a large sample sense, since rates are obtained for $n > N_0$, where N_0 itself might depend on d.

[2] So-called *self-similarity* conditions (roughly upper and lower-bounds of similar order on *smoothness*), which can be rather unnatural. Similarly restrictive, the earlier result Minsker (2012a) required the equivalence of L_{2,P_X} and L_{∞,P_X} distances between η and certain piecewise approximations to η (see Assumption 2 therein).

[3] A set of high confidence level (honesty) and of optimal size in terms of the unknown smoothness α (adaptivity).

case where the rate remains $n^{-\alpha(\beta+1)/(2\alpha+d)}$. Our lower-bound analysis suggests that $[d - (\alpha \wedge 1)\beta]_+$ plays the role of *degrees of freedom* in active learning - this is the case when $\alpha \leq 1, \beta \geq 0$ and in the case $\alpha \geq 1, \beta = 1$.

- For unrestricted $\mathbb{P}_X$, i.e., without the near uniform assumption, we prove that the minimax rate is of the form $n^{-\alpha(\beta+1)/(2\alpha+d)}$, showing a sharp difference between the regimes of uniform $\mathbb{P}_X$ and unrestricted $\mathbb{P}_X$. This difference mirrors the case of passive learning where the unrestricted $\mathbb{P}_X$ rate is of order $n^{-\alpha(\beta+1)/(2\alpha+d+\alpha\beta)}$. Again the key quantity in the rate-reduction from passive to active is the interaction term $\alpha\beta$.

In the case $\alpha < 1$ and $\mathbb{P}_X$ nearly uniform, we recover the rate $n^{-\alpha(\beta+1)/(2\alpha+[d-\alpha\beta]_+)}$ of Minsker (2012a,b) - but while avoiding the restrictive assumptions that are necessary therein to ensure that adaptive and honest confidence sets exist.

Algorithmic results. We present a generic strategy that avoids the need for honest confidence sets but is able to *adapt in an efficient way* to the unknown parameters α, β of the problem, simultaneously for all statistical regimes discussed above. Indeed our algorithm does not take the oracle values of α, β as parameters and yet achieves the oracle rate, over a large range of values of α, β (converging to any range of α with sufficiently large n). The main insight is a reduction to the case where α is known: iterating over $\alpha \approx 0$ to higher values, the procedure aggregates the estimates of a non-adaptive subroutine taking α as a parameter. This reduction is made possible by the nested structure of Hölder classes indexed by α: that is, η is α'-Hölder for any α' smaller than the true unknown α. Note that such nested class structure is also harnessed for adaptation in the passive setting as in Lepski and Spokoiny (1997), using techniques suited to passive sampling.

This reduction in active learning is perhaps of independent interest as it likely extends to any hierarchy of model classes. The reduction takes care of adaptivity to unknown α. What remains is to show that, for known α, there exists an efficient subroutine that adapts to unknown noise level β; fortunately, adaptivity to β comes for free once we have proper control of the bias and variance of local estimates of $\eta(x)$ (over a hierarchical partition of the feature space). Such control is easiest for $\alpha \leq 1$ and yields useful intuition towards handling the harder case $\alpha > 1$. Our final solution is a subroutine which, given α, actively labels the X space while requesting few Y values over a hierarchical space partition; it is computationally efficient and easy to implement.

Paper outline. We start in Section 2 with a detailed discussion of related work. We give the formal statistical setup in Section 3, followed by the main results and discussion in Sections 4 (main results, i.e., adaptive upper bounds and lower bounds). These results build on technical non-adaptive results presented in Sections 5. Section A contains all detailed proofs.

2. Related Work in Active Learning

Much of the theory in active learning covers a range of distributional assumptions which unfortunately are not always compatible or easy to compare with the present setting. We give an overview below of the current theory, and compare rates at the intersection of assumptions whenever feasible.

Parametric settings. Much of the current theory in active classification deals with the *parametric setting*. Such work is concerned with performance w.r.t. the best classifier over a fixed class $\mathcal{F} \equiv \{f : \mathbb{R}^d \mapsto \{0,1\}\}$ of small *complexity*, e.g., bounded VC dimension. It is well known that the passive rates in this case are of the form $n^{-1/2}$, i.e., have no dependence on d in the exponent; this is due to the relative small complexity of such $\mathcal{F}$, and corresponds[4] roughly to *infinite smoothness* in our case (indeed $n^{-1/2}$ is the limit of the nonparametric rates $n^{-\alpha/(2\alpha+d)}$ as $\alpha \to \infty$ and $\beta = 0$, i.e., no margin assumption).

The parametric theory has developed relatively fast, yielding much insight as to the relevant interaction between $\mathcal{F}$ and $\mathbb{P}_{X,Y}$. In particular, works such as Hanneke (2007); Dasgupta et al. (2007); Balcan et al. (2008, 2009); Beygelzimer et al. (2009) show that significant savings are possible over passive learning, provided the pair $(\mathcal{F}, \mathbb{P}_{X,Y})$ has bounded *Alexander capacity* (a.k.a. *disagreement-coefficient*, see Alexander (1987)). To be precise, the active rates are of the form[5] $\nu \cdot n^{-1/2} + \exp(-n^{1/2})$ where $\nu \doteq \inf_{f\in\mathcal{F}} \operatorname{err}(f)$; in other words the active rates behave like $\exp(-n^{1/2})$ when $\nu \approx 0$ (low noise), but otherwise are $\mathcal{O}(n^{-1/2})$ as in the passive case. More recently, Zhang and Chaudhuri (2014) shows similar rate regimes without requiring bounded disagreement coefficient.

Such rates are tight as shown by matching lower-bounds of Kääriäinen (2006), and Raginsky and Rakhlin (2011). This suggests that a refined parametrization of the noise regimes is needed to better capture the gains in active learning. The task is undertaken in the works of Hanneke (2009); Koltchinskii (2010) where the active rates are of the form $n^{-(\beta+1)/2}$, in terms of noise margin[6] β, and clearly show gains over known passive rates of the form $n^{-(\beta+1)/(\beta+2)}$. While this parametric setting is inherently different from ours, interestingly, our rates coincide at the intersection where $\mathbb{P}_X$ is unrestricted and we let $\alpha \to \infty$ (check that $\lim_{\alpha\to\infty} n^{-\alpha(\beta+1)/(2\alpha+d)} = n^{-(\beta+1)/2}$).

Nonparametric settings. Further results in Hanneke (2009) and Koltchinskii (2010) concern a setting where the class $\mathcal{F}$ is of larger complexity encoded in terms of *metric entropy*. The active rates in this case are of the form $n^{-(\beta+1)/(2+\rho\beta)}$, where ρ captures the complexity of $\mathcal{F}$. These rates are again better than the corresponding passive rate of $n^{-(\beta+1)/(2+\beta+\rho\beta)}$ shown earlier in Tsybakov (2004), but are only valid for classes with a bounded *disagreement coefficient*.

The complexity term ρ can be viewed as describing the richness of the Bayes decision boundary. This term becomes clear in the setting where the decision boundary is given by a $(d-1)$-dimensional curve of smoothness α' (to be interpreted as the graph of an α'-Hölder function $\mathbb{R}^{d-1} \mapsto \mathbb{R}$), in which case $\rho = (d-1)/\alpha'$ (as shown in Tsybakov (2004)). While it has been shown in Wang (2011) that under these assumptions the disagreement coefficient is unbounded and disagreement-based strategies lead to suboptimal rates, the earlier work of Castro and Nowak (2008) shows that active rates of the form $n^{-(\beta+1)/(2+\rho\beta)}$ are indeed tight in this nonparametric setting. Notice that the earlier parametric rates above correspond to $\rho = 0$, i.e., $\alpha' \to \infty$.

Unfortunately such active rates are hard to compare across settings, since boundary assumptions are inherently incompatible with smoothness assumptions on η: it is not hard to see that smooth η does not preclude complex boundary, neither does smooth boundary preclude complex η (as discussed in Audibert and Tsybakov (2007)). However, smoothness assumptions on η seem to be a richer setting that displays a variety of noise-regimes with different statistical rates, as shown here.

As discussed in the introduction, the closest work to ours is that of Minsker (2012a,b), as both works consider procedures that are efficient (unlike that of Koltchinskii (2010)[7]) and adaptive (unlike that of Castro and Nowak (2008)). However, our distinct algorithmic strategy yields interesting new insights on the effect of noise parameters under strictly broader statistical conditions.

[4] To compare across settings, we view $\mathcal{F}$ as the set of classifiers $\mathbb{I}\{\eta \geq 1/2\}$, where η is α-smooth.

[5] Omitting constants depending on the disagreement-coefficient.

[6] The rates are given in terms of a noise parameter $\kappa = (\beta+1)/\beta$ (see relation in Prop. 1 of Tsybakov (2004)).

[7] The procedure requires inefficient book-keeping over $\mathcal{F}$ as it discards functions with large error.

Other lines of work in Machine Learning are of a nonparametric nature given the estimators employed. The statistical aims are however different from ours. In particular Dasgupta and Hsu (2008); Urner et al. (2013); Kpotufe et al. (2015) are primarily concerned with the rates at which a fixed sample $\{X_i\}_1^n$ might be labeled, rather than in excess risk over the Bayes classifier. Interestingly, notions of smoothness and noise-margin (parametrized differently) also play important roles in such problems. In Kontorovich et al. (2016) on the other hand, the main concern is that of *sample-dependent* rates, i.e., rates that are given in terms of noise-characteristics of a random sample, rather than of the distribution as studied here.

It is important to note that, a recent procedure of Hanneke (2017), which is yet unpublished, concerns the same setting as ours, for the special case $\alpha \leq 1, \alpha\beta \leq d$ and uniform P_X, and achieves the minimax active rate of $n^{-\alpha(\beta+1)/(2\alpha+d-\alpha\beta)}$ without requiring adaptive honest confidence sets; instead the procedure follows insights similar to techniques presented in Kontorovich et al. (2016).

Finally, we remark that active learning is believed to be related to other sequential learning problems such as *bandits*, and *stochastic optimization*, and recent works such as Ramdas and Singh (2013) show that insights on noise regimes in active learning can cross over to such problems.

3. Preliminaries

3.1. The active learning setting

Let the feature-label pair (X, Y) have joint-distribution $\mathbb{P}_{X,Y}$, where the marginal distribution according to variable X is noted $\mathbb{P}_X$ and is supported on $[0,1]^d$, and where the random variable Y belongs to $\{0,1\}$. The conditional distribution of Y knowing $X = x$, which we denote $\mathbb{P}_{Y|X=x}$, is then fully characterized by the regression function

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

gWe extend the definition of η on $\mathbb{R}^d$ arbitrarily, so that we have $\eta : \mathbb{R}^d \mapsto [0,1]$ (although we are primarily concerned about its behavior on $[0,1]^d$). It is well known that the Bayes classifier $f^*(x) = \mathbf{1}\{\eta(x) \geq 1/2\}$ 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

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

Active sampling. At any point in time, the active learner can sample a label Y at any $x \in \mathbb{R}^d$ according to a Bernoulli random variable of parameter $\eta(x)$, i.e. according to the marginal distribution $P_{Y|X=x}$ if $x \in [0,1]^d$. The learner can request at most $n \in \mathbb{N}^*$ samples (i.e. its budget is n), and then returns a classifier $\widehat{f}_n : [0,1]^d \mapsto \{0,1\}$.

Our goal is therefore to design a sampling strategy that outputs a classifier $\widehat{f}_n$ whose excess risk $\mathcal{E}(\widehat{f}_n)$ is as small as possible, with high probability over the samples requested.

3.2. Assumptions and Definitions

We first define a hierarchical partitioning of $[0,1]^d$. This will come in handy in our subroutines.

Definition 1 *[Dyadic grid G_l, cells C, center x_C, and diameter r_l] We write G_l for the regular dyadic grid on the unit cube of mesh size 2^{-l}. It defines naturally a partition of the unit cube in 2^{ld}*

smaller cubes, or cells $C \in G_l$. They have volume 2^{-ld} and their edges are of length 2^{-l}. We have $[0,1]^d = \bigcup_{C \in G_l} C$ and $C \cap C' = \emptyset$ if $C \neq C'$, with $C, C' \in G_l^2$. We define x_C as the center of $C \in G_l$, i.e. the barycenter of C.
The diameter of the cell C is written :

$$r_l \doteq \max_{x,y \in C} |x-y|_2 = \sqrt{d}2^{-l}, \tag{2}$$

where $|z|_2$ is the Euclidean norm of z.

We now state the following assumption on $\mathbb{P}_X$.

Assumption 1 (Strong density) *There exists $c_1 > 0$ such that for all $l \geq 0$ and any cell C of G_l satisfying $\mathbb{P}_X(C_l) > 0$, we have:*

$$\mathbb{P}_X(C_l) \geq c_1 2^{-ld}.$$

This assumption allows us to lower bound the measure of a cell of the grid, and holds for instance when $\mathbb{P}_X$ is uniform or approximately so. This assumption is slightly weaker than the one in Minsker (2012a). We obtain results for both when Assumption 1 holds, and when it does not.

Definition 2 (Hölder smoothness) *For $\alpha > 0$ and $\lambda > 0$, we denote the Hölder class $\Sigma(\lambda, \alpha)$ of functions $g : \mathbb{R}^d \to [0,1]$ that are $\lfloor \alpha \rfloor$ times continuously differentiable, that are such that for any $j \in \mathbb{N}, j \leq \alpha$*

$$\sup_{x \in \mathbb{R}^d} \sum_{s:|s|=j} |D^s g(x)| \leq \lambda, \quad \textit{and,} \quad \sup_{x,y \in \mathbb{R}^d} \sum_{s:|s|=\lfloor \alpha \rfloor} \frac{|D^s g(x) - D^s g(y)|}{|x-y|_2^{\alpha - \lfloor \alpha \rfloor}} \leq \lambda,$$

where $D^s f$ is the classical mixed partial derivative with parameter s. Note that for $\alpha \leq 1$ and $\lambda \geq 1$, we simply require $\sup_{x,y \in \mathbb{R}^d} \frac{|g(y)-g(x)|}{|y-x|_2^{\alpha}} \leq \lambda$.

If a function is α-Hölder, then it is smooth and well approximated by polynoms of degree $\lfloor \alpha \rfloor$, but also by other approximation means, as e.g. Kernels.

Assumption 2 (Hölder smoothness of η) *η belongs to $\Sigma(\lambda, \alpha)$ with $\alpha > 0$ and $\lambda \geq 1$.*

We finally state our last assumption, which upper bounds the measure of the space where it is not easy to determine which class is best fitted.

Assumption 3 (Margin condition) *There exists nonnegative c_3, Δ_0, and β such that $\forall \Delta > 0$:*

$$\mathbb{P}_X(|\eta(X) - 1/2| < \Delta_0) = 0, \quad \textit{and,} \quad \mathbb{P}_X(|\eta(X) - 1/2| \leq \Delta_0 + \Delta) \leq c_3 \Delta^{\beta}.$$

These parameters cover many interesting cases, including $\Delta_0 = 0, \beta > 0$ (Tsybakov's noise condition) and $\Delta_0 > 0, \beta = 0$ (Massart's margin condition), which are common in the literature. This assumption allows us to bound the measure of regions close to the decision boundary (i.e. where η is close to $1/2$). The case $\Delta_0 > 0$ is linked to the *cluster assumption* in the semi-supervised learning literature (see e.g. Chapelle et al. (2003); Rigollet (2007)), and can model situations where $\text{supp}(\mathbb{P}_X)$ breaks up into components each admitting one dominant class (i.e. $|\eta - 1/2| \geq \Delta_0$ on each such component and η does not cross $1/2$ on $\text{supp}(\mathbb{P}_X)$).

Definition 3 *We denote by $\mathcal{P}(\alpha, \beta, \Delta_0) \doteq \mathcal{P}(\alpha, \beta, \Delta_0; \lambda, c_3)$ the set of classification problems $\mathbb{P}_{X,Y}$ characterized by $(\eta, \mathbb{P}_X)$ that are such that Assumptions 2 and 3 are satisfied with parameters $\alpha > 0, \beta \geq 0, \Delta_0 \geq 0$, and some fixed $\lambda \geq 1, c_3 > 0$. Moreover, we denote $\mathcal{P}^*(\alpha, \beta, \Delta_0)$ the subset of $\mathcal{P}(\alpha, \beta, \Delta_0)$ such that $\mathbb{P}_X$ satisfies Assumption 1 (strong density).*

We fix in the rest of the paper $c_3 > 0$ and $\lambda \geq 1$. These parameters will be discussed in Section 4.2.

4. Adaptive Results

We start with a detailed presentation of our main adaptive strategy, Algorithm 1.

Algorithm 1 Adapting to unknown smoothness α

```
Input: $n$, $\delta$, $\lambda$, and a black-box Subroutine
Initialization: $s_0^0 = s_0^1 = \emptyset$
for $i = 1, ..., \lfloor \log(n) \rfloor^3$ do
    Let $n_0 = \frac{n}{\lfloor \log(n) \rfloor^3}$, $\delta_0 = \frac{\delta}{\lfloor \log(n) \rfloor^3}$, and $\alpha_i = \frac{i}{\lfloor \log(n) \rfloor^2}$
    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: $S^0 = s^0_{\lfloor \log(n) \rfloor^3}, S^1 = s^1_{\lfloor \log(n) \rfloor^3}$ and classifier $\hat{f}_n = \mathbf{1}\{S^1\}$
```

Algorithm 1 aggregates the label estimates of a black-box (non-adaptive) Subroutine over increasing guesses α_i of the unknown smoothness parameter α. Algorithm 1 takes as parameters n, δ, λ, and the black-box Subroutine, and outputs a classifier $\hat{f}_n$. Here n is the sampling budget, δ is the desired level of confidence of the algorithm, λ is such that η is (λ, α)-Hölder for some unknown α; in practice λ is also unknown, but any upper-bound is sufficient, e.g. $\log n$ for n sufficiently large.

In each phase $i \in \{1, 2, \ldots, \lfloor \log(n) \rfloor^3\}$, the black-box Subroutine takes four parameters: a sampling budget n_0, a confidence level δ_0, and smoothness parameters α_i, λ. It then returns two disjoint subsets of $[0, 1]^d$, $S_i^y, y \in \{0, 1\}$. The set S_i^0 corresponds to all $x \in [0, 1]^d$ that are labeled 0 by the Subroutine (in phase i), and S_i^1 corresponds to the label 1. The remaining space $[0, 1]^d \setminus S_i^1 \cup S_i^0$ corresponds to a region that the Subroutine could not confidently label.

Algorithm 1 calls the Subroutine $\lfloor \log(n) \rfloor^3$ times, for increasing values of α_i on the grid $\{\lfloor \log(n) \rfloor^{-2}, 2\lfloor \log(n) \rfloor^{-2}, ..., \lfloor \log(n) \rfloor\})$, and collects the sets S_i^y that it aggregates into s_i^y. For n sufficiently large, this grid contains the unknown α parameter to be adapted to.

The main intuition behind the procedure relies on the nestedness of Hölder classes: if η is α-Hölder for some unknown α, then it is α_i-Hölder for $\alpha_i \leq \alpha$. Thus, suppose the Subroutine returns *correct* labels S_i^y whenever η is α_i-Hölder; then for any $\alpha_i \leq \alpha$ the aggregated labels remain correct. When $\alpha_i > \alpha$, the error cannot be higher than the error in earlier phases since the aggregation never overwrites correct labels. In other words, the excess risk of Algorithm 1 is at most the error due to the highest phase s.t. $\alpha_i \leq \alpha$. We therefore just need the Subroutine to be correct in an *optimal* way formalized below.

Definition 4 **((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* $\mathbb{P}_{X,Y}$ *(characterized by* $(\eta, \mathbb{P}_X)$*) if, with probability larger than* $1 - 8\delta$ *over at most* n *label requests:*

$$\Big\{x \in [0,1]^d : \eta(x) - 1/2 > \Delta\Big\} \subset S^1, \text{ and } \qquad \Big\{x \in [0,1]^d : 1/2 - \eta(x) > \Delta\Big\} \subset S^0.$$

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

$$S^1 \subset \Big\{x \in [0,1]^d : \eta(x) - 1/2 > 0\Big\}, \text{ and } \qquad S^0 \subset \Big\{x \in [0,1]^d : \eta(x) - 1/2 < 0\Big\},$$

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

4.1. Main Results

We now present our main results, which are high-probability bounds on the risk of the classifier output by Algorithm 1, under different noise regimes. Our upper-bounds build on the following simple proposition, the intuition of which was detailed above.

Proposition 1 (Correctness of aggregation) *Let* $n \in \mathbb{N}^*$ *and* $1 > \delta > 0$. *Let* $\delta_0 = \delta/(\lfloor\log(n)\rfloor^3)$ *and* $n_0 = n/(\lfloor\log(n)\rfloor^3)$ *as in Algorithm 1. Fix* $\beta \geq 0$, $\Delta_0 \geq 0$. *Suppose that, for any* $\alpha > 0$, *the Subroutine in Algorithm 1 is* $(\delta_0, \Delta_\alpha, n_0)$*-correct for any* $\mathbb{P}_{X,Y} \in \mathcal{P}^*(\alpha, \beta, \Delta_0)$*, where* $0 \leq \Delta_\alpha$ *depends on* n, δ *and the class* $\mathcal{P}^*(\alpha, \beta, \Delta_0)$.

Fix $\alpha \in [\lfloor\log(n)\rfloor^{-2}, \lfloor\log(n)\rfloor]$, *and let* $\alpha_i = i/\lfloor\log(n)\rfloor^2$ *for* $i \in \{1, \ldots, \lfloor\log(n)\rfloor^3\}$. *Then Algorithm 1 is* **weakly** $(\delta_0, \Delta_{\alpha_i}, n_0)$*-correct for any* $\mathbb{P}_{X,Y} \in \mathcal{P}^*(\alpha, \beta, \Delta_0)$ *for the largest* i *such that* $\alpha_i \leq \alpha$.

The same holds true for $\mathcal{P}(\alpha, \beta, \Delta_0)$ *in place of* $\mathcal{P}^*(\alpha, \beta, \Delta_0)$.

Remark 4.1 To see why the proposition is useful, suppose for instance that our problem belongs to $\mathcal{P}^*(\alpha, \beta, 0)$, and Algorithm 1 happens to be weakly (δ_0, Δ, n_0)-correct on this problem for some $\Delta \doteq \Delta(n_0, \delta_0, \alpha, \beta)$. Then, by definition of correctness, the returned classifier $\hat{f}_n$ agrees with the Bayes classifier f on the set $\{x : |\eta(x) - 1/2| > \Delta\}$; that is, its excess error only happens on the set $\{x : |\eta(x) - 1/2| \leq \Delta\}$. Therefore by Equation (1), with probability larger than $1 - \delta_0$

$$\mathcal{E}(\hat{f}_n) \leq 2\Delta \cdot \mathbb{P}_X\left(\{x : |\eta(x) - 1/2| \leq \Delta\}\right) \leq 2c_3\Delta^{1+\beta}.$$

In other words, we just need to show the existence of a Subroutine which is (δ_0, Δ, n_0)-correct for any class $\mathcal{P}^*(\alpha, \beta, \Delta_0)$ (or respectively $\mathcal{P}(\alpha, \beta, \Delta_0)$) with $\Delta \doteq \Delta(n_0, \delta_0, \alpha, \beta, \Delta_0)$ of appropriate order over ranges of α, β, Δ_0. The adaptive results on the next sections are derived in this manner. In particular, we will show that Algorithm 2 of Section 5 is a *correct* such Subroutine.

Our results show that the excess risk rates in the active setting are strictly faster than in the passive setting (except for $\beta = 0$, i.e., no noise condition), in both cases i.e. when $\mathbb{P}_X$ is nearly uniform on its support (Assumption 1), and when it is fully unrestricted. These two cases are presented in the next two sections.

4.1.1. ADAPTIVE RATES FOR $\mathcal{P}^*(\alpha, \beta, \Delta_0)$

We start with results for the class $\mathcal{P}^*(\alpha, \beta, \Delta_0)$, i.e. under the *strong density* condition which encodes the usual assumption in previous work that the marginal $\mathbb{P}_X$ is nearly uniform.

Theorem 1 (Adaptive upper-bounds) *Let $n \in \mathbb{N}^*$ and $1 > \delta > 0$. Assume that $\mathbb{P}_{X,Y} \in \mathcal{P}^*(\alpha, \beta, \Delta_0)$ with $\left(\frac{3d}{\log(n)}\right)^{1/3} \leq \alpha \leq \lfloor \log(n) \rfloor$.*

Algorithm 1, with input parameters $(n, \delta, \lambda, \mathit{Algorithm}\ 2)$, outputs a classifier $\widehat{f}_n$ satisfying the following, with probability at least $1 - 8\delta$:

- *For any $\Delta_0 \geq 0$,*

$$\mathcal{E}(\widehat{f}_n) \leq C \left(\frac{\lambda^{(\frac{d}{\alpha \wedge 1} \vee \beta)} \log^3(n) \log(\frac{\lambda n}{\delta})}{n} \right)^{\frac{\alpha(\beta+1)}{2\alpha + [d - (\alpha \wedge 1)\beta]_+}},$$

  where the constant $C > 0$ does not depend on n, δ, λ.

- *If $\Delta_0 > 0$, then $\mathcal{E}(\widehat{f}_n) = 0$ whenever the budget satisfies*

$$\frac{n}{\lfloor \log(n) \rfloor^3} > C \log\left(\frac{\lambda n}{\delta}\right) \cdot \left(\frac{\lambda^{(\frac{d}{\alpha \wedge 1} \vee \beta)}}{\Delta_0} \right)^{\frac{2\alpha + [d - (\alpha \wedge 1)\beta]_+}{\alpha}}$$

  where $C > 0$ does not depend on n, δ, λ.

The above theorem is proved, following Remark 4.1, by showing that Algorithm 3 is *correct* for problems in $\mathcal{P}^*(\alpha, \beta, \Delta_0)$ with some $\Delta = \mathcal{O}(n^{-\alpha/(2\alpha + [d - (\alpha \wedge 1)\beta]_+})$; for $\Delta_0 > 0$, correctness is obtained for $\Delta \leq \Delta_0$, provided sufficiently large budget n. See Theorem 6.

The rate of Theorem 1 matches (up to logarithmic factors) the minimax lower-bound for this class of problems with $\alpha > 0, \beta \geq 0$ such that $\alpha\beta \leq d$ obtained in Minsker (2012a), which we recall hereunder for completeness.

Theorem 2 (Lower-bound: Theorem 7 in Minsker (2012a)) *Let $\alpha > 0, \beta \geq 0$ such that $\alpha\beta \leq d$ and assume that c_3, λ are large enough. For n large enough, any (possibly active) strategy that samples at most n labels and returns a classifier $\widehat{f}_n$ satisfies :*

$$\sup_{\mathbb{P}_{X,Y} \in \mathcal{P}^*(\alpha, \beta, 0)} \mathbb{E}_{\mathbb{P}_{X,Y}}[\mathcal{E}_{\mathbb{P}_{X,Y}}(\widehat{f}_n)] \geq C n^{-\frac{2\alpha}{2\alpha + d - \alpha\beta}},$$

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

However, the above lower-bound turns out not to be tight for $\alpha > 1$. We now present a novel minimax lower-bound that complements the above, and which is always tighter for $\alpha > 1, \beta = 1$. To the best of our knowledge, it is the first lower bound that highlights the phase transition in the active learning setting for $\alpha > 1$ which was conjectured in Minsker (2012a).

Theorem 3 (Lower-bound) *Let $\alpha > 0$, $\beta = 1$, and assume that c_3, λ are large enough. For n large enough, any (possibly active) strategy that samples at most n labels and returns a classifier $\widehat{f}_n$ satisfies:*

$$\sup_{\mathbb{P}_{X,Y} \in \mathcal{P}^*(\alpha,1,0)} \mathbb{E}_{\mathbb{P}_{X,Y}}[\mathcal{E}_{\mathbb{P}_{X,Y}}(\widehat{f}_n)] \geq C n^{-\frac{2\alpha}{2\alpha+d-1}},$$

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

The proof of this new lower-bound is given in Section A.3 of the Appendix.

Remark 4.2 Under the strong density assumption, the rate is improved from $n^{-\alpha(\beta+1)/(2\alpha+d)}$ to $n^{-\alpha(\beta+1)/(2\alpha+[d-(\alpha\wedge 1)\beta]_+)}$. This implies that fast rates (i.e. faster than $n^{-1/2}$) are reachable for $\alpha\beta > d/(2+(\alpha\wedge 1)^{-1})$, improving from $\alpha\beta > d/2$ in the passive learning setting. This rate matches (up to logarithmic factors) the lower-bound in Minsker (2012a) for $\alpha \leq 1$.

It also improves on the results in Minsker (2012a), as we require strictly weaker assumptions (see Assumption 2 in Minsker (2012a), which in light of the examples given is rather strong). In the important case $\alpha > 1$, our results match the rate conjectured in Minsker (2012a), up to logarithmic factors. The conjectured rates of Minsker (2012a) turns out to be tight, as our lower-bound shows for the case $\beta = 1$, i.e. no better upper-bound is possible over all β. This highlights that there is indeed a phase transition happening (at least when $\beta = 1$) when we go from the case $\alpha \leq 1$ to the case $\alpha \geq 1$. Our lower-bound leaves open the possibility of even richer transitions over regimes of the β parameter.

Our lower-bound analysis of Section A.3 shows that, at least for $\beta = 1$, the quantity $d - \beta$ acts like the *degrees of freedom* of the problem: we can make η change fast in at least $d - \beta$ directions, and this is sufficient to make the problem difficult.

4.1.2. ADAPTIVE RATES FOR $\mathcal{P}(\alpha, \beta, \Delta_0)$

We now exhibit a theorem very similar to Theorem 1, but that holds for more general classes, as we do not impose regularity assumptions on the marginal $\mathbb{P}_X$, which is thus *unrestricted*.

Theorem 4 (Upper-bound) *Let $n \in \mathbb{N}^*$ and $1 > \delta > 0$. Assume that $\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \beta, \Delta_0)$ with $\frac{1}{\lfloor \log(n) \rfloor} \leq \alpha \leq \lfloor \log(n) \rfloor$.*

Algorithm 1, with input parameters $(n, \delta, \lambda, Algorithm\ 2)$, outputs a classifier $\widehat{f}_n$ satisfying the following, with probability at least $1 - 8\delta$:

- *For any Δ_0:*

$$\mathcal{E}(\widehat{f}_n) \leq C\lambda^{\frac{d(\beta+1)}{2\alpha+d}} \left(\frac{\log^3(n)\log(\frac{\lambda n}{\delta})}{n}\right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},$$

  where $C > 0$ does not depend on n, δ, λ.

- *If $\Delta_0 > 0$, then $\mathcal{E}(\widehat{f}_n) = 0$ whenever the budget satisfies*

$$\frac{n}{\lfloor \log^3(n) \rfloor} > C\lambda^{d/\alpha} \log\left(\frac{\lambda n}{\delta}\right)\left(\frac{1}{\Delta_0}\right)^{\frac{2\alpha+d}{\alpha}}$$

  where $C > 0$ does not depend on n, δ, λ.

The above theorem is proved, following Remark 4.1, by showing that Algorithm 3 is *correct* for problems in $\mathcal{P}(\alpha, \beta, \Delta_0)$ with some $\Delta = \mathcal{O}(n^{-\alpha/(2\alpha+d)})$; for $\Delta_0 > 0$, correctness is obtained for $\Delta \leq \Delta_0$, provided sufficiently large budget n. See Theorem 7.

We complement this result with a novel lower-bound for this class of problems, which shows that the result in Theorem 4 is tight up to logarithmic factors.

Theorem 5 (Lower-bound) *Let $\alpha > 0, \beta \geq 0$ and assume that c_3, λ are large enough. For n large enough, any (possibly active) strategy that samples at most n labels and returns a classifier $\widehat{f}_n$ satisfies:*

$$\sup_{\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha,\beta,0)} \mathbb{E}_{\mathbb{P}_{X,Y}}[\mathcal{E}_{\mathbb{P}_{X,Y}}(\widehat{f}_n)] \geq C n^{-\frac{\alpha(1+\beta)}{2\alpha+d}},$$

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

The proof of this last theorem is given in Section A.4 of the Appendix.

Remark 4.4 The unrestricted $\mathbb{P}_X$ case treated in this section is analogous to the *mild* density assumptions studied in Audibert and Tsybakov (2007) in the passive setting. Our results imply that even under these weaker assumptions, the active setting brings an improvement in the rate - from $n^{-\alpha(\beta+1)/(2\alpha+d+\alpha\beta)}$ to $n^{-\alpha(\beta+1)/(2\alpha+d)}$. The rate improvement is possible since an active procedure can save in labels by focusing all samplings to regions where η is close to $1/2$. However, this might not be possible in passive learning since the density in such regions can be arbitrarily low and thus yield too few training samples. To better appreciate the improvement in rates, notice that the passive rates are never faster than n^{-1}, while in the active setting, we can reach super fast rates (i.e. faster than n^{-1}) as soon as $\alpha\beta > d$. In fact, this rate is similar to the minimax optimal rate in the passive setting under the strong density assumption: in some sense the active setting mirrors the strong density assumption, given the ability of the learner to sample everywhere.

4.2. General Remarks

Adaptivity to the unknown parameters. An important feature of Algorithm 1 is that it is *adaptive* to the parameters α, β, Δ_0 from Assumptions 2 and 3 - i.e. it does not take these parameters as inputs and yet has smaller excess risk than the minimax optimal excess risk rate over all classes $\mathcal{P}(\alpha, \beta, \Delta_0)$ (respectively $\mathcal{P}^*(\alpha, \beta, \Delta_0)$ if Assumption 1 holds) to which the problem belongs to. A key point in the construction of Algorithm 1 is that it makes use of the nested nature of the models. A different strategy could have been to use a cross-validation scheme to select one of the classifiers output by the different runs of Algorithm 2, however such a strategy would not allow fast rates, as the cross-validation error might dominate the rate. Instead, taking advantage of the nested smoothness classes, we can aggregate our classifiers such that the resulting classifier is in agreement with all the classifiers that are optimal for bigger classes - this idea is related to the construction in the totally different passive setting Lepski and Spokoiny (1997). This aggregation method is an important feature of our algorithm, as it bypasses the calculation of disagreement sets or other quantities that can be computationally intractable, such as optimizing over entire sets of functions as in Hanneke (2009); Koltchinskii (2010). It also allows us to remove a key restriction on the class of problems in Minsker (2012a) - see Assumption 2 therein required for the construction of *honest and adaptive confidence sets*. Our algorithm moreover adapts to the parameter c_3 of Assumptions 3, but takes as parameter λ of Assumption 2. However, it is possible to use in the algorithm an upper bound on the parameter λ - as e.g. $\log(n)$ for n large enough - and to only worsen the excess risk bound by a λ at a bounded power - e.g. poly $\log(n)$.

Extended Settings. Note that our results can readily be extended to the multi-class setting (see Dinh et al. (2015) for the multi-class analogous of Audibert and Tsybakov (2007) in the passive setting) through a small but necessary refinement of the aggregation method (one has to keep track of eliminated classes i.e. classes deemed impossible for a certain region of the space by bigger models). It is also possible to modify Assumption 1 such that the box-counting dimension of the support of $\mathbb{P}_X$ is $d' < d$ (if for example $\mathbb{P}_X$ is supported on a manifold of dimension d' embedded in $[0,1]^d$), and we would obtain similar results where d is replaced by d', effectively adapting to that smaller dimension.

5. Non-Adaptive Subroutine

In this section, we construct an algorithm that is optimal over a given smoothness class $\Sigma(\lambda, \alpha)$ - and that uses the knowledge of λ, α. This algorithm is non-adaptive, as is often the case in the continuum-armed bandit literature that assumes knowledge of a semi-metric in order to optimize (i.e. maximize or minimize) the sum of rewards gathered by an agent receiving noisy observations of a function (Auer et al. (2007), Kleinberg et al. (2008), Cope (2009), Bubeck et al. (2011)).

5.1. Description of the Subroutine

Algorithm 2 Non-adaptive Subroutine

```
Input: $n$, $\delta$, $\alpha$, $\lambda$
Initialisation: $t = 2^d t_{1,\alpha\wedge 1}$, $l = 1$, $\mathcal{A}_1 \doteq G_1$ (active space), $\forall l' > 1, \mathcal{A}_{l'} \doteq \emptyset$, $S^0 = S^1 \doteq \emptyset$
while $t + |\mathcal{A}_l| \cdot t_{l,\alpha} \leq n$ do
  for each active cell $C \in \mathcal{A}_l$ do
    Request $t_{l,\alpha\wedge 1}$ samples $(\tilde{Y}_{C,i})_{i \leq t_{l,\alpha\wedge 1}}$ at the center $x_C$ of $C$
    if $\left\{ |\widehat{\eta}(x_C) - 1/2| \leq B_{l,\alpha} \right\}$ then
      $\mathcal{A}_{l+1} = \mathcal{A}_{l+1} \cup \{C' \in G_{l+1} : C' \subset C\}$          // keep all children $C'$ of $C$ active
    else
      Let $y \doteq \mathbf{1}\{\widehat{\eta}(x_C) \geq 1/2\}$
      $S^y = S^y \cup C$                                  // label the cell as class $y$
    end if
  end for
  Increase depth to $l = l + 1$, and set $t \doteq t + |\mathcal{A}_l| \cdot t_{l,\alpha\wedge 1}$
end while
Set $L = l - 1$
if $\alpha > 1$ then
  Run Algorithm 3 on last partition $\mathcal{A}_L$
end if
Output: $S^y$ for $y \in \{0, 1\}$, and $\hat{f}_{n,\alpha} = \mathbf{1}\{S^1\}$
```

We first introduce an algorithm that takes λ, α as parameters, and refines its exploration of the space to focus on zones where the classification problem is the most difficult (i.e. where η is close to the $1/2$ level set). It does so by iteratively refining a partition of the space (based on a dyadic tree), and using a simple plug-in rule to label cells. At a given depth l, the algorithm samples the center x_C of the *active cells* $C \in \mathcal{A}_l$ a fixed number of times $t_{l,\alpha\wedge 1}$ with:

$$t_{l,\alpha} = \begin{cases} \frac{\log(1/\delta_{l,\alpha})}{2b_{l,\alpha}^2} \text{ if } \alpha \leq 1 \\ 4^{2d+1}(\alpha+1)^{2d}\frac{\log(1/\delta_{l,\alpha})}{b_{l,\alpha}^2} \text{ if } \alpha > 1, \end{cases}$$

where $b_{l,\alpha} = \lambda d^{(\alpha\wedge 1)/2}2^{-l\alpha}$ and $\delta_{l,\alpha} = \delta 2^{-l(d+1)(\alpha\vee 1)}$, and collects the labels $(\tilde{Y}_{C,i})_{i\leq t_{l,\alpha\wedge 1}}$. The algorithm then compares an estimate $\widehat{\eta}(x_C)$ of $\eta(x_C)$ with $1/2$. The estimate is simply the sample-average of Y-values at x_C, i.e.:

$$\widehat{\eta}(x_C) = t_{l,\alpha\wedge 1}^{-1}\sum_{i=1}^{t_{l,\alpha\wedge 1}} \tilde{Y}_{C,i}.$$

If $|\widehat{\eta}(x_C) - 1/2|$ is sufficiently large with respect to

$$B_{l,\alpha} = 2\Big[\sqrt{\frac{\log(1/\delta_{l,\alpha\wedge 1})}{2t_{l,\alpha\wedge 1}}} + b_{l,\alpha\wedge 1}\Big],$$

which is the sum of a bias and a deviation term, the cell is labeled (i.e. added to S^1 or S^0) as the best empirical class, i.e. as

$$\mathbf{1}\{\widehat{\eta}(x_C) \geq 1/2\},$$

and we refer to that process as *labeling*. If the gap is too small then the partition needs to be refined, and the cell is split into smaller cubes. All these cells are then the *active cells* at depth $l+1$. The algorithm stops refining the partition of the space when a given constraint on the used budget is saturated, namely when the used budget t plus $t_{l,\alpha}.|\mathcal{A}_l|$ is larger than n - this happens at depth L.

If $\alpha \geq 1$, we need to consider higher order estimators in active cells - we make use of smoothing Kernels to take advantage of the higher smoothness to estimate η more precisely. This last step is described in Algorithm 3. For any $l \geq 1$ and any cell $C \in G_l$, we write $\tilde{C}$ for the *inflated cell* C, such that

$$\tilde{C} = \{x \in \mathbb{R}^d : \inf_{z\in C}\sup_{i\leq d}|x^{(i)} - z^{(i)}| \leq 2^{-l}\},$$

where $x^{(i)}, z^{(i)}$ are the ith coordinates of respectively x, z.

A number $t_{L,\alpha}$ of samples $(X_{C,i}, Y_{C,i})_{C\in\mathcal{A}_L, i\leq t_{L,\alpha}}$ is collected uniformly at random in each inflated cell $\tilde{C}$ corresponding to any $C \in \mathcal{A}_L$. For any $\alpha > 0$, let $\tilde{k}_\alpha$ the one-dimensional convolution Kernel of order $\lfloor\alpha\rfloor + 1$ based on the Legendre polynomial, defined in the proof of Proposition 4.1.6 in Giné and Nickl (2015). Consider the d-dimensional corresponding isotropic product Kernel defined for any $z \in \mathbb{R}^d$ as :

$$K_\alpha(z) = \prod_{i=1}^{d}\tilde{k}_\alpha(z^{(i)}).$$

The Subroutine then updates S^0 and S^1 in the active regions of $\mathcal{A}_L$ using the Kernel estimator

$$\hat{\eta}_C(x) = \frac{1}{t_{l,\alpha}}\sum_{i\leq t_{l,\alpha}} K_\alpha((x - X_{C,i})2^l)Y_{C,i}.$$

Finally (both when $\alpha \leq 1$ and $\alpha > 1$) the algorithm returns the sets S^0, S^1 of labeled cells in classes respectively 0 or 1 and uses them to build the classifier $\hat{f}_n$ - the cells that are still active receive an arbitrary label (here 0).

Algorithm 3 Procedure for smoothness $\alpha > 1$

```
for each cell $C \in \mathcal{A}_L$ do
  Sample uniformly $t_{L,\alpha}$ points $(X_{C,i}, Y_{C,i})_{i \leq t_{L,\alpha}}$ on $\tilde{C}$
  for each cell $C' \in G_{\lfloor L\alpha \rfloor}$ such that $C' \subset C$ do
    Set
      $$\widehat{\eta}_C(x_{C'}) = \frac{1}{t_{L,\alpha}} \sum_{i \leq t_{L,\alpha}} K_\alpha((x_{C'} - X_{C,i})2^L) Y_{C,i}.$$
    Set $S^0 = S^0 \cup C'$, if $\widehat{\eta}_C(x_{C'}) - 1/2 < 4^{d+1}\lambda 2^{-\alpha L}$
    Set $S^1 = S^1 \cup C'$, if $\widehat{\eta}_C(x_{C'}) - 1/2 > 4^{d+1}\lambda 2^{-\alpha L}$
  end for
end for
```

5.2. Non-Adaptive Results

The first result is for the class $\mathcal{P}^*(\lambda, \alpha, \beta, \Delta_0)$, in particular under the *strong density* assumption.

Theorem 6 *Algorithm 2 run on a problem in $\mathcal{P}^*(\lambda, \alpha, \beta, \Delta_0)$ with input parameters $n, \delta, \alpha, \lambda$ is $(\delta, \Delta^*_{n,\delta,\alpha,\lambda}, n)-$correct, with*

$$\Delta^*_{n,\delta,\alpha,\lambda} = \begin{cases} 12\sqrt{d}\Big(\frac{c_7 \lambda^{(\frac{d}{\alpha} \vee \beta)} \log\left(\frac{2d\lambda^2 n}{\delta}\right)}{(2\alpha + [d-\alpha\beta]_+)\alpha\ n}\Big)^{\frac{\alpha}{2\alpha+[d-\alpha\beta]_+}} \textit{for} \ \ \alpha \leq 1 \\ 4^{d+2}2^\alpha\Big(\frac{c_8 \lambda^{(d\vee\beta)} \log(\frac{2d\lambda^2 n}{\delta})}{n}\Big)^{\frac{\alpha}{2\alpha+[d-\beta]_+}} \ \ \textit{otherwise,} \end{cases}$$

with $c_7 = 2(d+1)c_5$, $c_8 = 4^{2d+1}(\alpha+1)^{2\alpha}(d+1)c_5$ and $c_5 = 2^{(\alpha\wedge 1)\beta}\max(\frac{c_3}{c_1}8^\beta, 1)$, where c_1 and c_3 are the constants involved in Assumption 1 and 3 respectively.

The proof of this theorem is in Section A.1 of the Appendix.
An important case to consider is that if $\Delta_0 > 0$, then the excess risk of the classifier output by Algorithm 2 is nil with probability $1 - 8\delta$ as soon as $\Delta^*_{n,\delta,\alpha,\lambda} < \Delta_0$. Inverting the bound on $\Delta^*_{n,\delta,\alpha,\lambda}$ for n yields a sufficient condition on the budget, that we made clear in Theorem 1.

We now exhibit another theorem, very similar to Theorem 6, but that holds for more general classes, as we do not impose regularity assumptions on the density.

Theorem 7 *Algorithm 2 run on a problem in $\mathcal{P}(\lambda, \alpha, \beta, \Delta_0)$ with input parameters $n, \delta, \alpha, \lambda$ is $(\delta, \Delta_{n,\delta,\alpha,\lambda}, n)-$correct, with*

$$\Delta_{n,\delta,\alpha,\lambda} = \begin{cases} 12\sqrt{d}\lambda^{d/(2\alpha+d)}\Big(\frac{2(d+1)\log\left(\frac{2d\lambda^2 n}{\delta}\right)}{(2\alpha+d)\alpha\ n}\Big)^{\frac{\alpha}{2\alpha+d}} \textit{for} \ \ \alpha \leq 1 \\ 4^{d+2}2^\alpha\lambda^{d/(2\alpha+d)}\Big(\frac{4^{2d+1}(\alpha+1)^{2d}(d+1)\log(\frac{2d\lambda^2 n}{\delta})}{n}\Big)^{\frac{\alpha}{2\alpha+d}} \ \ \textit{otherwise.} \end{cases}$$

The proof of this theorem is in Section A.1 of the Appendix.

These results show that Algorithm 2 can be used by Algorithm 1 for any problem $\mathbb{P}_{X,Y} \in \mathcal{P}^*(\alpha, \beta, \Delta_0)$ (respectively $\mathbb{P}_{X,Y} \in \mathcal{P}(\alpha, \beta, \Delta_0)$), as it is $(\delta, \Delta^*_{n,\delta,\alpha,\lambda}, n)-$correct (respectively $(\delta, \Delta_{n,\delta,\alpha,\lambda}, n)-$correct).

5.3. Remarks on Non-Adaptive Procedures

Optimism in front of uncertainty. The main principle behind our algorithm is that of optimism in face of uncertainty, as we label regions thanks to an optimistic lower-bound on the gap between η and its $1/2$ level set, borrowing from well understood ideas in the bandit literature (see Auer et al. (2002), Bubeck et al. (2012)), which translate naturally to the continuous-armed bandit problem (see Auer et al. (2007); Kleinberg et al. (2008)). This allows the algorithm to prune regions of the space for which it is confident that they do not intersect the $1/2$ level set, in order to focus on regions harder to classify (w.r.t. $1/2$), naturally adapting to the margin conditions.

Hierarchical partitioning. Our algorithm proceeds by keeping a hierarchical partition of the space, zooming in on regions that are not yet classified with respect to $1/2$. This kind of construction is related to the ones in Bubeck et al. (2011); Munos (2011) that target the very different setting of *optimization of a function*. It is also related to the strategies exposed in Perchet et al. (2013), which tackles the *contextual* bandit problem in the setting where $\alpha \leq 1$ - in this setting the learner does not actively explore the space but receives random features.

Conclusion

In this work, we presented a new active strategy that is adaptive to various regimes of noise conditions. Our results capture interesting rate transitions under more general conditions than previously known. Some interesting open questions remain, including the possibility of even richer rate-transitions under a more refined parametrization of the problem.

Acknowledgement The work of A. Carpentier and A. Locatelli is supported by the DFG's Emmy Noether grant MuSyAD (CA 1488/1-1).

References

Oracle inequalities in empirical risk minimization and sparse recovery problems.

K.S. Alexander. Rates of growth and sample moduli for weithed empirical processes indexed by sets. Probability Theory and Related Fields, 75(3):379–423, 1987.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. The Annals of statistics, 35(2):608–633, 2007.

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.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In International Conference on Computational Learning Theory, pages 454–468. Springer, 2007.

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. COLT, 2008.

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

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. ICML, 2009.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. X-armed bandits. Journal of Machine Learning Research, 12(May):1655–1695, 2011.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends® in Machine Learning, 5(1):1–122, 2012.

Adam D Bull and Richard Nickl. Adaptive confidence sets in l_2. Probability Theory and Related Fields, 156(3-4):889–919, 2013.

T Tony Cai, Mark G Low, et al. Adaptive confidence balls. The Annals of Statistics, 34(1):202–228, 2006.

Alexandra Carpentier. Honest and adaptive confidence sets in l_p. Electronic Journal of Statistics, 7: 2875–2923, 2013.

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

Olivier Chapelle, Jason Weston, and Bernhard Schölkopf. Cluster kernels for semi-supervised learning. In Advances in Neural Information Processing Systems, pages 601–608, 2003.

Eric Cope. Regret and convergence bounds for immediate-reward reinforcement learning with continuous action spaces. IEEE Transactions on Automatic Control, 54(6):1243–1253, 2009.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. NIPS, 2007.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In Proceedings of the 25th international conference on Machine learning, pages 208–215. ACM, 2008.

Vu Dinh, Lam Si Tung Ho, Nguyen Viet Cuong, Duy Nguyen, and Binh T Nguyen. Learning from non-iid data: Fast rates for the one-vs-all multiclass plug-in classifiers. In International Conference on Theory and Applications of Models of Computation, pages 375–387. Springer, 2015.

Christopher Genovese and Larry Wasserman. Adaptive confidence bands. The Annals of Statistics, pages 875–905, 2008.

Evarist Giné and Richard Nickl. Mathematical foundations of infinite-dimensional statistical models, volume 40. Cambridge University Press, 2015.

S. Hanneke. A bound on the label complexity of agnostic active learning. ICML, 2007.

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

Steve Hanneke. Nonparametric active learning, part 1: Smooth regression functions. Unpublished, 2017.

Marc Hoffmann and Richard Nickl. On adaptive inference and confidence bands. The Annals of Statistics, pages 2383–2409, 2011.

Matti Kääriäinen. Active learning in the non-realizable case. In International Conference on Algorithmic Learning Theory, pages 63–77. Springer, 2006.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In Proceedings of the fortieth annual ACM symposium on Theory of computing, pages 681–690. ACM, 2008.

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

Aryeh Kontorovich, Sivan Sabato, and Ruth Urner. Active nearest-neighbor learning in metric spaces. In Advances in Neural Information Processing Systems, pages 856–864, 2016.

Samory Kpotufe, Ruth Urner, and Shai Ben-David. Hierarchical label queries with data-dependent partitions. In COLT, pages 1176–1189, 2015.

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

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

Stanislav Minsker. Non-asymptotic bounds for prediction problems and density estimation. PhD thesis, Georgia Institute of Technology, 2012b.

Rémi Munos. Optimistic Optimization of Deterministic Functions without the Knowledge of its Smoothness. In Advances in Neural Information Processing Systems, 2011.

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

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. NIPS, 2011.

Aaditya Ramdas and Aarti Singh. Algorithmic connections between active learning and stochastic convex optimization. In ALT, volume 8139, pages 339–353. Springer, 2013.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. Journal of Machine Learning Research, 8(Jul):1369–1392, 2007.

James Robins, Aad Van Der Vaart, et al. Adaptive nonparametric confidence sets. The Annals of Statistics, 34(1):229–253, 2006.

Alexandre Tsybakov. Introduction to nonparametric estimation. 2009.

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

Ruth Urner, Sharon Wullf, and Shai Ben-David. Plal: Cluster-based active learning. In Proceedings of the Conference on Learning Theory (COLT), 2013.

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

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In Advances in Neural Information Processing Systems, pages 442–450, 2014.

Appendix A. Proofs of the theoretical results

A.1. Proof of Theorem 6 and Theorem 7

A.1.1. PROOF OF THEOREM 6

Let us write in this proof in order to simplify the notations

$$t_l = t_{l,\alpha\wedge 1}, \quad b_l = b_{l,\alpha\wedge 1}, \quad \delta_l = \delta_{l,\alpha\vee 1}, \quad B_l = B_{l,\alpha\wedge 1} \quad \text{and} \quad N_l = |\mathcal{A}_l|.$$

We will now show that on a certain event, the algorithm makes no mistake up to a certain depth L, and that the error is controlled beyond that depth.

Step 1: A favorable event.

Consider a cell C of depth l. We define the event:

$$\xi_{C,l} = \Big\{|t_l^{-1}\sum_{u=1}^{t_l}\mathbf{1}(\tilde{Y}_{C,i}=1) - \eta(x_C)| \le \sqrt{\frac{\log(1/\delta_l)}{2t_l}}\Big\},$$

where the $(\tilde{Y}_{C,i})_{i\le t_l}$ are samples collected in C at point x_C if C if the algorithm samples in cell C. We remind that

$$\widehat{\eta}(x_C) = t_l^{-1}\sum_{i=1}^{t_l}\mathbf{1}(\tilde{Y}_{C,i}=1).$$

We consider the following event ξ:

$$\xi = \Big\{\bigcap_{l\in\mathbb{N}^*, C\in G_l}\xi_{C,l}\Big\}.$$

Lemma 1 *We have*

$$\mathbb{P}(\xi) \ge 1-4\delta.$$

Moreover on ξ

$$|\widehat{\eta}(x_C) - \eta(x_C)| \le b_l. \tag{3}$$

Step 2: No mistakes on labeled cells.

For $l \in \mathbb{N}^*$, let $C \in G_l$ and write

$$\widehat{k}^*_C = \mathbf{1}\{\widehat{\eta}(x_C) \ge 1/2\} \text{ and let us write, } k^*_C \doteq \mathbf{1}\{\eta(x_C)\ge 1/2\}.$$

Lemma 2 *We have that on ξ,*

$$\forall y\in\{0,1\}, \forall C\in S^y, \forall x\in C, \quad \mathbf{1}\{\eta(x)\ge 1/2\} = y. \tag{4}$$

This implies that:

$$S^1 \subset \{x : \eta(x) - 1/2 > 0\} \textit{ and, } S^0 \subset \{x : \eta(x) - 1/2 < 0\}. \tag{5}$$

Step 3: Maximum gap with respect to $1/2$ for all active cells.

Now we will consider a cell C that is split and added to $\mathcal{A}_{l+1}$ at depth $l \in \mathbb{N}^*$ by the algorithm. As C is split and added to $\mathcal{A}_{l+1}$, we have by definition of the algorithm and on ξ using Equation (3)

$$|\eta(x_C) - 1/2| - b_l \leq |\widehat{\eta}(x_C) - 1/2| \quad \leq 4b_l,$$

which implies $|\eta(x_C) - 1/2| \leq 5b_l$. Using Equation (13), this implies that on ξ for any C that will be split and added to $\mathcal{A}_{l+1}$ and for any $x \in C$

$$|\eta(x) - 1/2| \leq 6b_l \doteq \Delta_l. \tag{6}$$

Step 4: Bound on the number of active cells.

Set for $\Delta \geq 0$

$$\Omega_\Delta = \Big\{x \in [0,1]^d : |\eta(x) - 1/2| \leq \Delta\Big\},$$

and let for $l \in \mathbb{N}^*$, $N_l(\Delta)$ be the number of cells $C \in G_l$ such that $C \subset \Omega_\Delta$.

Lemma 3 *We have on ξ*

$$\begin{aligned} N_{l+1} &\leq \frac{c_3}{c_1}[\Delta_l - \Delta_0]_+^\beta r_{l+1}^{-d} \\ &\leq c_5\lambda^\beta r_{l+1}^{-[d-(\alpha\wedge 1)\beta]_+}\mathbf{1}_{\Delta_l > \Delta_0}, \end{aligned} \tag{7}$$

Step 5: A minimum depth.

Lemma 4 *We have on ξ the following results on L.*

- *Case a) : If $\alpha \leq 1$: It holds that*

$$L \geq \frac{1}{2\alpha + [d-\alpha\beta]_+}\log_2\Big(\frac{(2\alpha + [d-\alpha\beta]_+)2\alpha n}{c_7\lambda^{\beta-2}\log\big(\frac{2d\lambda^2 n}{\delta}\big)}\Big) - 1, \tag{8}$$

  with $c_7 = 2c_5(d+1)$, or the algorithm stops before reaching depth L and $\mathcal{E}(\widehat{f}_n) = 0$.

- *Case b) : If $\alpha > 1$:*

$$L \geq \frac{1}{2\alpha + [d-\beta]_+}\log_2\big(\frac{n}{c_8\lambda^{\beta-2}\log(\frac{2d\lambda^2 n}{\delta})}\big) - 1, \tag{9}$$

  where $c_8 = c_5 4^{2d+1}(\alpha+1)^{2d}(d+1)$, or the algorithm stops before reaching depth L and $\mathcal{E}(\widehat{f}_n) = 0$.

Step 6 : Conclusion.
From this point on, we write S^0, S^1 for the sets that Algorithm 2 outputs at the end (so the sets at the end of the algorithm).

We write the following lemma.

Lemma 5 *If $S^1 \cap S^0 = \emptyset$ and if for some $\Delta \geq 0$ we have on some event ξ'*

$$\{x \in [0,1]^d : \eta(x) - \Delta_L \geq 1/2\} \subset S^1, \;\; \text{and} \;\; \{x \in [0,1]^d : \eta(x) + \Delta_L \leq 1/2\} \subset S^0,$$

then on ξ' it holds that

$$\sup_{x \in [0,1]^d : \hat{f}_{n,\alpha} \neq f^*(x)} |\eta(x) - 1/2| \leq \Delta_L, \;\; \text{and} \;\; \mathbb{P}_X(\hat{f}_{n,\alpha} \neq f^*) \leq c_3 \Delta_L^\beta \mathbf{1}\{\Delta \geq \Delta_0\},$$

and

$$\mathcal{E}(\widehat{f}_{n,\alpha}) \leq c_3 \Delta_L^{1+\beta} \mathbf{1}\{\Delta_L \geq \Delta_0\}.$$

Proof The first conclusion is a direct consequence of the lemma's assumption, the second conclusions follows directly from the lemma's assumption and Assumption 3, and the third conclusion follows as

$$\mathcal{E}(\widehat{f}_{n,\alpha}) \leq \mathbb{P}_X(\hat{f}_{n,\alpha}) \neq f^*) \sup_{x \in [0,1]^d} |\hat{f}_{n,\alpha}(x) - f^*(x)|.$$

■

CASE a) : $\alpha \leq 1$.

Note first that $S^1 \cap S^0 = \emptyset$ by definition of the algorithm. By Equation (6) and Equation (4), we know that on ξ (and so with probability larger than $1 - 4\delta$)

$$\{x \in [0,1]^d : \eta(x) - \Delta_L \geq 1/2\} \subset S^1, \;\; \text{and,} \;\; \{x \in [0,1]^d : \eta(x) + \Delta_L \leq 1/2\} \subset S^0, \qquad (10)$$

where

$$\begin{aligned}
\Delta_L &\leq 6\lambda d^{\alpha/2} 2^\alpha \Big(\frac{c_7 \lambda^{\beta-2} \log\big(\frac{2d\lambda^2 n}{\delta}\big)}{(2\alpha + [d-\alpha\beta]_+) 2\alpha n} \Big)^{\alpha/(2\alpha+[d-\alpha\beta]_+)} \\
&\leq 12\lambda d^{\alpha/2} \Big(\frac{c_7 \lambda^{\beta-2} \log\big(\frac{2d\lambda^2 n}{\delta}\big)}{(2\alpha + [d-\alpha\beta]_+) 2\alpha n} \Big)^{\alpha/(2\alpha+[d-\alpha\beta]_+)} \\
&\leq 12\sqrt{d} \Big(\frac{c_7 \lambda^{(\frac{d}{\alpha} \vee \beta)} \log\big(\frac{2d\lambda^2 n}{\delta}\big)}{(2\alpha + [d-\alpha\beta]_+) 2\alpha n} \Big)^{\alpha/(2\alpha+[d-\alpha\beta]_+)}
\end{aligned}$$

by Equation (8). This implies the first part of Theorem 6 for $\alpha \leq 1$.
So by Lemma 5, we have on ξ (and so with probability larger than $1 - 4\delta$)

$$\begin{aligned}
\sup_{x \in [0,1]^d : \widehat{f}_{n,\alpha} \neq f^*(x)} |\eta(x) - 1/2| &\leq \Delta_L \\
&\leq 12\sqrt{d} \Big(\frac{c_7 \lambda^{(\frac{d}{\alpha} \vee \beta)} \log\big(\frac{2d\lambda^2 n}{\delta}\big)}{(2\alpha + [d-\alpha\beta]_+) 2\alpha n} \Big)^{\alpha/(2\alpha+[d-\alpha\beta]_+)},
\end{aligned}$$

and also

$$\begin{aligned}
\mathbb{P}_X(\hat{f}_{n,\alpha} \neq f^*(x)) &\leq c_3 \Delta_L^\beta \mathbf{1}(\Delta_L \geq \Delta_0) \\
&\leq c_3 12^\beta \sqrt{d} \Big(\frac{c_7 \lambda^{(\frac{d}{\alpha} \vee \beta)} \log\big(\frac{2d\lambda^2 n}{\delta}\big)}{(2\alpha + [d-\alpha\beta]_+) 2\alpha n} \Big)^{\alpha\beta/(2\alpha+[d-\alpha\beta]_+)}
\end{aligned}$$

and also that

$$\begin{aligned}\mathcal{E}(\widehat{f}_{n,\alpha}) &\leq c_3\Delta_L^{\beta+1}\mathbf{1}(\Delta_L \geq \Delta_0)\\ &\leq c_3 12^{\beta+1}\sqrt{d}\Big(\frac{c_7\lambda^{(\frac{d}{\alpha}\vee\beta)}\log\left(\frac{2d\lambda^2 n}{\delta}\right)}{(2\alpha+[d-\alpha\beta]_+)2\alpha n}\Big)^{\alpha(\beta+1)/(2\alpha+[d-\alpha\beta]_+)}.\end{aligned}$$

CASE b) : $\alpha > 1$.
Denote $\widehat{\eta}_C$ the estimator built in the second phase of the algorithm, described in Lemma 6.

Let us write $(X_{C,i}, Y_{C,i})_{u\leq t_{l,\alpha}}$ for the (not necessarily observed) samples that would be collected in $\tilde{C}$ if cell $C \in \mathcal{A}_L$. For any $x \in C$ and any cell C, we write

$$\widehat{\eta}_C(x) = \frac{1}{t_{l,\alpha}}\sum_{i\leq t_{l,\alpha}} K_\alpha((x - X_{C,i})2^l)Y_{C,i}.$$

Note that $\hat{\eta}_C$ is computed by the algorithm for any $C \in \mathcal{A}_L$ (and $\hat{\eta}$ is $1/2$ everywhere else).

The following proposition holds.

Proposition 2 *Let* $l > 0$, $C \in G_l$ *and assume that* $\eta \in \Sigma(\lambda,\alpha)$*. It holds for* $x \in C$ *that with probability larger than* $1-\delta$

$$|\widehat{\eta}_C(x) - \eta(x)| \leq 4^d\lambda 2^{-l\alpha} + 2^{d+2}(2\alpha+2)^d\sqrt{\frac{\log(1/\delta)}{t_{l,\alpha}}}.$$

Let

$$\xi' = \Big\{\forall l \geq 1, \forall C \in G_{\lfloor l\alpha\rfloor}, |\widehat{\eta}_C(x_C) - \eta(x_C)| \leq \lambda\sqrt{d}2^{-l\alpha}\Big\}.$$

Since $\delta_{l,\alpha} = \delta 2^{-l\alpha(d+1)}$, it holds by Proposition 2 and an union bound that this event holds with probability at least $1-4\delta$. By a union bound, the event $(\xi \cap \xi')$ thus holds with probability at least $1-8\delta$.

By Proposition 2, and proceeding as in Step 3, we can bound on ξ' the maximum gap of the cells that are not classified i.e. cells C such that $C \cap (S^0 \cup S^1) = \emptyset$. Recall that if $\alpha > 1$ then by Assumption 2, η is λ-Lipschitz. For cells of side length $2^{-\lfloor L\alpha\rfloor}$, this yields for any $x \in C$ such that $|\widehat{\eta}_C(x_C) - 1/2| \leq 4^{d+1}\lambda 2^{-L\alpha}$:

$$\begin{aligned}|\eta(x) - 1/2| &\leq (4^{d+3/2} + 3\sqrt{d})\lambda 2^{-L\alpha}\\ &\leq 4^{d+2}\lambda 2^{-L\alpha}\end{aligned}$$

On the other hand, for $x \in C$ such that $|\widehat{\eta}(x_C) - 1/2| > 4^{d+1}\lambda 2^{-L\alpha}$, we have:

$$|\eta(x) - 1/2| > 4^d\lambda 2^{-L\alpha}, \tag{11}$$

which implies that:

$$\{x \in [0,1]^d : \eta(x) - \Delta_L \geq 1/2\} \subset S^1 \subset \{x : \eta(x) - 1/2 > 0\}$$

and

$$\{x \in [0,1]^d : \eta(x) + \Delta_L \leq 1/2\} \subset S^0 \subset \{x : \eta(x) - 1/2 < 0\}.$$

with:

$$\Delta_L \leq 4^{d+2} 2^{\alpha} \Big(\frac{c_8 \lambda^{(d \vee \beta)} \log(\frac{2d\lambda^2 n}{\delta})}{n} \Big)^{\alpha/(2\alpha+[d-\beta]_+)}, \tag{12}$$

where we lower bound L using Equation (19). We conclude the proof by using Lemma 5 as in the case $\alpha \leq 1$, and the result holds with probability at least $1 - 8\delta$.

A.1.2. Proof of Theorem 7

The proof of this result only differs from the proof of Theorem 6 in Step 4, Equation (7), where we can no longer use the lower bound on the density to upper bound the number of active cells, and instead we have to use the naive bound 2^{-ld} at depth l such that all cells can potentially be active. The rest of the technical derivations is similar to the case $\beta = 0$ in the proof of Theorem 6.

A.1.3. Proofs of the technical lemmas and propositions stated in the proof of Theorem 6 and 7

Proof [Proof of Lemma 1] From Hoeffding's inequality, we know that $\mathbb{P}(\xi_{C,l}) \geq 1 - 2\delta_l$.

We now consider

$$\xi = \Big\{ \bigcap_{l \in \mathbb{N}^*, C \in G_l} \xi_{C,l} \Big\},$$

the intersection of events such that for all depths l and any cell $C \in G_l$, the previous event holds true. Note that at depth l there are 2^{ld} such events. A simple union bound yields $\mathbb{P}(\xi) \geq 1 - \sum_l 2^{ld}\delta_l \geq 1 - 4\delta$ as we have set $\delta_l = \delta 2^{-l(d+1)}$.

We define $b_l = \lambda d^{(\alpha\wedge 1)/2} 2^{-l(\alpha \wedge 1)}$ for any $l \in \mathbb{N}^*$. By Assumption 2, it is such that for any $x, y \in C$, where $C \in G_l$, we have:

$$|\eta(x) - \eta(y)| \leq b_l. \tag{13}$$

On the event ξ, for any $l \in \mathbb{N}^*$, as we have set $t_l = \frac{\log(1/\delta_l)}{2b_l^2}$, plugging this in the bound yields that at time t_l, we have for each cell $C \in G_l$:

$$|\widehat{\eta}(x_C) - \eta(x_C)| \leq b_l.$$

■

Proof [Proof of Lemma 2] Using Equations (13) and (3), we have:

$$4b_l < \widehat{\eta}_{\widehat{k}^*_C}(x_C) - 1/2 < \eta_{\widehat{k}^*_C}(x_C) + b_l - 1/2,$$

which implies that $\eta_{\widehat{k}^*_C}(x_C) - 1/2 > 3b_l > 0$. So necessarily by definition of k^*_C, we have $k^*_C = \widehat{k}^*_C$.

Now using the smoothness assumption, we have for any $x \in C$:

$$|\eta(x) - \eta(x_C)| \leq \lambda d^{(\alpha \wedge 1)/2} |x - x_C|_2^{\alpha \wedge 1} \leq b_l.$$

Assume now without loss of generality that $\widehat{k}^*_C = 1$. We have by the previous paragraph that $\widehat{k}^*_C = k^*_C = 1$ and that $\eta(x_C) - 1/2 > 2b_l$. So for $x \in C$, $\eta_{k^*_C}(x) - 1/2 > 0$, so k^*_C is the best class in the entire cell C and the labeling $\widehat{k}^*_C = k^*_C$ is in agreement with that of the Bayes classifier *on the entire cell*, bringing no excess risk on the cell. In summary we have that on ξ,

$$\forall y \in \{0,1\}, \forall C \in S^y, \forall x \in C, \qquad \mathbf{1}\{\eta(x) \geq 1/2\} = y.$$

This implies that:

$$S^1 \subset \{x : \eta(x) - 1/2 > 0\} \text{ and, } S^0 \subset \{x : \eta(x) - 1/2 < 0\}.$$

■

Proof [Proof of Lemma 3] Since by Assumption 3 we have $\mathbb{P}_X(\Omega) \leq c_3(\Delta - \Delta_0)^\beta \mathbf{1}\{\Delta \geq \Delta_0\}$, we have by Assumption 1 that

$$N_l(\Delta) \leq \frac{c_3}{c_1}(\Delta - \Delta_0)^\beta r_l^{-d} \mathbf{1}\{\Delta \geq \Delta_0\}. \tag{14}$$

Let us write L for the depth of the active cells at the end of the algorithm. The previous equation implies with Equation (6) that on ξ, for $l \leq L$, the number of cells in $\mathcal{A}_l$ is bounded as Equation (14) brings

$$\begin{aligned} N_{l+1} &\leq N_{l+1}(\Delta_l) \leq \frac{c_3}{c_1}[\Delta_l - \Delta_0]_+^\beta r_{l+1}^{-d} \\ &\leq \frac{c_3}{c_1} 8^\beta \lambda^\beta 2^{(\alpha\wedge 1)\beta} r_{l+1}^{(\alpha \wedge 1)\beta - d} \mathbf{1}_{\Delta_l > \Delta_0} \leq c_5 \lambda^\beta r_{l+1}^{-[d - (\alpha\wedge 1)\beta]_+} \mathbf{1}_{\Delta_l > \Delta_0}, \end{aligned}$$

where N_{l+1} is the number of active cells at the beginning of the round of depth $(l+1)$ and $[a]_+ = \max(0, a)$ and $c_5 = 2^{(\alpha \wedge 1)\beta} \max(\frac{c_3}{c_1} 8^\beta, 1)$. This formula is valid for $L - 1 \geq l \geq 0$. ■

Proof [Proof of Lemma 4]
CASE a): $\alpha \leq 1$.
At each depth $1 \leq l \leq L$, we sample these active cells $t_l = \frac{\log(1/\delta_l)}{2b_l^2}$ times. Let us first consider the case $\Delta_0 = 0$. We will upper-bound the total number of samples required by the algorithm to reach depth L. We know by Equation (7) that on ξ:

$$\begin{aligned} \sum_{l=1}^{L} N_l t_l + N_L t_L &\leq 2\sum_{l=1}^{L} (c_5 \lambda^\beta r_l^{-[d-\alpha\beta]_+}) \frac{\log(1/\delta_l)}{2\lambda^2 r_l^{2\alpha}} \\ &\leq 2c_5 \lambda^{\beta-2} \log(1/\delta_L) \sum_{l=1}^{L} r_l^{-(2\alpha + [d-\alpha\beta]_+)} \\ &\leq 2c_5 d^{-(2\alpha+d-\alpha\beta)/2} \lambda^{\beta-2} \log(1/\delta_L) \frac{2^{L(2\alpha+[d-\alpha\beta]_+)}}{2^{2\alpha+[d-\alpha\beta]_+} - 1} \\ &\leq \frac{4c_5}{d^{(2\alpha+d-\alpha\beta)/2}} \lambda^{\beta-2} \log(1/\delta_L) \frac{2^{L(2\alpha+[d-\alpha\beta]_+)}}{2\alpha + [d-\alpha\beta]_+}, \end{aligned}$$

as $2^a - 1 \geq a/2$ for any $a \in \mathbb{R}^+$. This implies that on ξ

$$\sum_{l=1}^{L} N_l t_l + N_L t_L \quad \leq \quad 4c_5 \lambda^{\beta-2} \log(1/\delta_L) \frac{2^{L(2\alpha+[d-\alpha\beta]_+)}}{2\alpha + [d-\alpha\beta]_+}. \tag{15}$$

We will now bound L by above naively, as t_L itself has to be smaller than n (otherwise, if there is a single active cell - which is the minimum number of active cells - the budget is not sufficient). This yields:

$$\frac{\log(1/\delta_L)}{2\lambda^2 r_L^{2\alpha}} \leq n,$$

which yields immediately, using $\delta_L < \delta \leq \mathrm{e}^{-1}$:

$$L \leq \frac{1}{2\alpha} \log_2 \big(2d\lambda^2 n\big).$$

We can now bound $\log(1/\delta_L)$:

$$\begin{aligned} \log(1/\delta_L) = \log(2^{L(d+1)}/\delta)) \quad &\leq \quad \frac{d+1}{2\alpha} \log \big(2d\lambda^2 n\big) + \log(1/\delta) \\ &\leq \quad \frac{d+1}{2\alpha} \log \big(\frac{2d\lambda^2 n}{\delta}\big). \end{aligned} \tag{16}$$

Combining equations (16) and (15), this implies that on ξ the budget is sufficient to reach the depth

$$L \geq \Big\lfloor \frac{1}{2\alpha + [d-\alpha\beta]_+} \log_2 \Big(\frac{(2\alpha + [d-\alpha\beta]_+) 2\alpha n}{c_7 \lambda^{\beta-2} \log\big(\frac{2d\lambda^2 n}{\delta}\big)} \Big) \Big\rfloor,$$

with $c_7 = 2c_5(d+1)$, or the algorithm stops before reaching the depth L with $S^1 \cup S^0 = [0,1]^d$, and the excess risk is 0.

CASE b): $\alpha > 1$.
We will proceed similarly as in the previous case. We have set $t_{l,\alpha} = 4^{2(d+1)}(\alpha+1)^{2d} \frac{\log(1/\delta_{l,\alpha})}{b_{l,\alpha}^2}$ with $b_{l,\alpha} = \lambda\sqrt{d} 2^{-l\alpha}$ and $\delta_{l,\alpha} = \delta 2^{-l\alpha(d+1)}$. By construction of the algorithm, L is the biggest integer such that $\sum_{l=1}^{L} N_l t_l + N_L t_{L,\alpha} \leq n$. We now bound this sum by above:

$$\begin{aligned} \sum_{l=1}^{L} N_l t_l + N_L t_{L,\alpha} \quad &\leq \quad \sum_{l=1}^{L} (c_5 \lambda^\beta r_l^{-[d-\beta]_+}) \frac{\log(1/\delta_l)}{2\lambda^2 r_l^2} + (4^{2d+1}(\alpha+1)^{2d} c_5 \lambda^\beta r_L^{-[d-\beta]_+}) \frac{\log(1/\delta_{L,\alpha})}{\lambda^2 d 2^{2L\alpha}} \\ &\leq \quad c_5 \lambda^{\beta-2} d^{-\frac{2+[d-\beta]_+}{2}} (\log(1/\delta_L) 2^{L(2+[d-\beta]_+)} + 4^{2d+1}(\alpha+1)^{2d} \log(1/\delta_{L,\alpha}) 2^{L(2\alpha+[d-\beta]_+)} \\ &\leq \quad 2c_5 \lambda^{\beta-2} 4^{2d+1} (\alpha+1)^{2d} \log(1/\delta_{L,\alpha}) 2^{L(2\alpha+[d-\beta]_+)} \end{aligned} \tag{17}$$

As in the previous case, we can upper bound L by remarking that $t_{L,\alpha}$ has to be smaller than the total budget n, which yields:

$$L \leq \frac{1}{2\alpha} \log_2(2d\lambda^2 n).$$

In turn, this allows to bound $\log(1/\delta_{L,\alpha})$:

$$\log(1/\delta_{L,\alpha}) = \log(2^{\alpha L(d+1)}/\delta) \quad \leq \quad \frac{d+1}{2} \log\big(\frac{2d\lambda^2 n}{\delta}\big) \tag{18}$$

Now combining Equations (17) and (18), it follows that on ξ, the budget is sufficient to reach a depth L such that:

$$L \geq \Big\lfloor \frac{1}{2\alpha + [d-\beta]_+} \log_2 \big(\frac{n}{c_8 \lambda^{\beta-2} \log(\frac{2d\lambda^2 n}{\delta})}\big) \Big\rfloor,$$

where $c_8 = c_5 4^{2d+1}(\alpha+1)^{2d}(d+1)$, or this depth is not reached as the algorithm stops with $S^1 \cup S^0 = [0,1]^d$ and the excess risk is 0.

■

Proof [Proof of Proposition 2]

The following Lemma holds regarding approximation properties of the Kernel we defined, see Giné and Nickl (2015).

Lemma 6 (Properties of the Legendre polynomial product Kernel K) *It holds that :*

- *K_α is non-zero only on $[-1,1]^d$.*
- *K_α is bounded in absolute value by $(2\alpha+2)^d$*
- *For any $h>0$ and any $\mathbb{P}_X$-measurable $f:\mathbb{R}^d \to [0,1]$,*

$$\sup_{x\in\mathbb{R}^d} |K_{\alpha,h}(f)(x) - f(x)| \leq 4^d \lambda h^\alpha, \quad \textit{where} \quad K_{\alpha,h}(f)(x) = \frac{1}{h^d}\int_{u\in\mathbb{R}^d} K_\alpha(\frac{x-u}{h}) f(u)du.$$

Proof The first and second properties follow immediately by definition of the Legendre polynomial Kernel $\tilde{k}_\alpha$ (see the proof of Proposition 4.1.6 from Giné and Nickl (2015)). The last property follows from the second result in Proposition 4.3.33 in Giné and Nickl (2015), which applies as Condition 4.1.4 in Giné and Nickl (2015) holds for $\tilde{k}_\alpha$ (see Proposition 4.1.6 from Giné and Nickl (2015) and its proof). ■

We bound separately the bias and stochastic deviations of our estimator.

Bias : We first have for any $x \in x_C + [-h,h]^d$

$$\begin{aligned}\mathbb{E}\hat{\eta}_C(x) &= \mathbb{E}\Big[2^d K((x-X_i)2^l)\eta(X_i) | X_i \text{ uniform on } \tilde{C}\Big]\\ &= 2^{ld}\int K((x-u)2^l)\eta(u)du,\end{aligned}$$

since X_i is uniform on $\tilde{C}$, and $x \in C$, and $K(\frac{x-.}{h})$ is 0 everywhere outside $\prod_i [x_i - 2^{-l}, x_i + 2^{-l}]$ (by Lemma 6). So by Lemma 6 we know that

$$|K_{2^{-l}}(\eta_C)(x) - \eta(x)| \leq 4^d \lambda 2^{-l\alpha}.$$

Deviation : Consider $Z_i = K((x - X_i)2^l)Y_i = K((x - X_i)2^l)f(X_i) + K((x - X_i)2^l)\epsilon_i$. Since by Lemma 6 $|K| \leq (2\alpha + 2)^d$, $\sup_x |\eta(x)| \leq 1$ and $|\epsilon_i| \leq 1$, we have by Hoeffding's inequality that with probability larger than $1 - \delta$:

$$|\mathbb{E}\hat{\eta}_C(x) - \hat{\eta}_C(x)| \leq 2^{d+2}(2\alpha + 2)^d \sqrt{\frac{\log(1/\delta)}{t_{l,\alpha}}}.$$

This concludes the proof by summing the two terms. ■

A.2. Proof of Proposition 1, Theorem 1 and 4

Set

$$n_0 = \frac{n}{\lfloor \log(n) \rfloor^3}, \quad \delta_0 = \frac{\delta}{\lfloor \log(n) \rfloor^3}, \quad \text{and} \quad \alpha_i = \frac{i}{\lfloor \log(n) \rfloor^2}.$$

A.2.1. PROOF OF PROPOSITION 1

In Algorithm 1, the Subroutine is launched $\lfloor \log(n) \rfloor^3$ times on $\lfloor \log(n) \rfloor^3$ 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 $1 \leq i \leq \lfloor \log(n) \rfloor^3$ such that $\alpha_i \leq \alpha$.

Since the Subroutine is strongly $(\delta_0, \Delta_\alpha, n_0)$-correct for any $\alpha \in [\lfloor \log(n) \rfloor^{-2}, \lfloor \log(n) \rfloor]$, it holds by Definition 4 that for any $i \leq i^*$, with probability larger than $1 - \delta_0$

$$\Big\{x \in [0,1]^d : \eta(x) - 1/2 \geq \Delta_{\alpha_i}\Big\} \subset S_i^1 \subset \Big\{x \in [0,1]^d : \eta(x) - 1/2 > 0\Big\}$$

and

$$\Big\{x \in [0,1]^d : \eta(x) - 1/2 \leq -\Delta_{\alpha_i}\Big\} \subset S_i^0 \subset \Big\{x \in [0,1]^d : \eta(x) - 1/2 < 0\Big\}.$$

So by an union bound we know that with probability larger than $1 - \lfloor \log(n) \rfloor^3 \delta_0 = 1 - \delta$, 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 $y \in \{0,1\}$

$$\bigcup_{i \leq i^*} S_i^y \subset s_{\lfloor \log(n) \rfloor^3}^y \quad \text{and} \quad s_{\lfloor \log(n) \rfloor^3}^y \cap s_{\lfloor \log(n) \rfloor^3}^{1-y} = \emptyset.$$

This finishes the proof of Proposition 1.

A.2.2. PROOF OF THEOREM 1, 4

The previous equation and Theorem 6 imply that with probability larger than $1-8\delta$

$$S^y_{i^*} \subset s^y_{\lfloor \log(n) \rfloor^3} \quad \text{and} \quad s^y_{\lfloor \log(n) \rfloor^3} \cap s^{1-y}_{\lfloor \log(n) \rfloor^3} = \emptyset.$$

So from Theorem 6, and Lemma 5, we have that with probability larger than $1-8\delta$

$$\mathcal{E}(\widehat{f}_n) \leq c_3 \Delta_{\alpha_{i^*}}^{\beta+1} \mathbf{1}(\Delta_{\alpha_{i^*}} \geq \Delta_0).$$

By definition of α_{i^*}, we know that it is such that:

$$\alpha - \frac{1}{\log^2(n)} \leq \alpha_{i^*} \leq \alpha. \tag{19}$$

In the setting of Theorem 1 for $\alpha \leq 1$ and $\alpha > \max(\sqrt{\frac{d}{2\log(n)}}, \big(\frac{3d}{\log(n)}\big)^{1/3})$, this yields for the exponent if $\alpha_{i^*}\beta \leq d$:

$$-\frac{\alpha_{i^*}(1+\beta)}{2\alpha_{i^*} + d - \alpha_{i^*}\beta} \quad \leq \quad -\frac{\alpha(1+\beta)}{2\alpha + [d-\alpha\beta]_+} + \frac{(1+\beta)(2\alpha+d)}{\log^2(n)(2\alpha + [d-\alpha\beta]_+)^2}.$$

The result follows by remarking that:

$$n^{\frac{(1+\beta)(2\alpha+d)}{\log^2(n)(2\alpha+d-\alpha\beta)^2}} = \exp\Big(\frac{(1+\beta)(2\alpha+d)}{\log(n)(2\alpha+[d-\alpha\beta]_+)^2}\Big),$$

and thus the extra additional term in the rate only brings at most a constant multiplicative factor, as the choice of $\alpha > \big(\frac{3d}{\log(n)}\big)^{1/3}$ allows us to upper-bound the quantity inside the exponential, using $\alpha - \log^{-3}(n) > \alpha/2$:

$$\frac{(1+\beta)(2\alpha+d)}{\log(n)(2\alpha+[d-\alpha\beta]_+)^2} \leq \frac{3d}{\log(n)\alpha^3} \leq 1.$$

In the case $\alpha > 1$ and $\beta < d$, first notice that $\alpha_{i^*} \geq 1$, as $\alpha_{\lfloor \log(n) \rfloor^2} = 1 < \alpha$. Thus, the rate can be rewritten:

$$-\frac{\alpha_{i^*}(1+\beta)}{2\alpha_{i^*} + [d-\beta]_+} \leq -\frac{\alpha(1+\beta)}{2\alpha + [d-\beta]_+} + \frac{1+\beta}{\log^2(n)(2\alpha+[d-\beta]_+)},$$

and the result follows.

In the case $(\alpha_{i^*} \wedge 1)\beta > d$, we immediately get $-\frac{1+\beta}{2}$, which is the desired rate.

For Theorem 4, we have instead:

$$-\frac{\alpha_{i^*}(1+\beta)}{2\alpha_{i^*}+d} \leq -\frac{\alpha_{i^*}(1+\beta)}{2\alpha+d} \leq -\frac{\alpha(1+\beta)}{2\alpha+d} + \frac{1+\beta}{\log^2(n)(2\alpha+d)},$$

which yields the desired rate.

The second part of the theorems is obtained by inverting the condition $\Delta_{\alpha_{i^*}} < \Delta_0$ for $\Delta_0 > 0$.

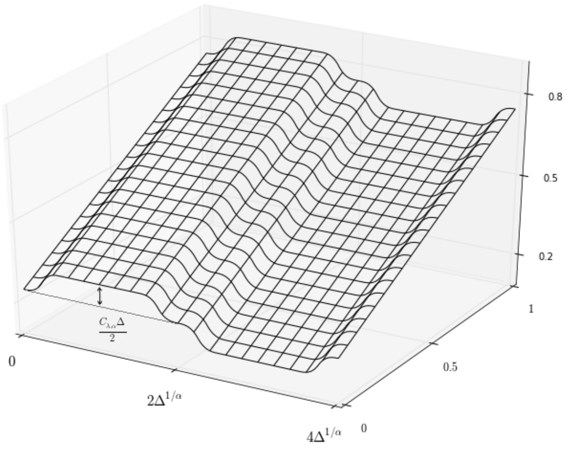

Figure 1: *Lower-bound construction of $\eta(x)$ illustrated for $d = 2$. The function changes slowly (linearly) in one direction, but can change fast – at most α smooth, in $d-\beta$ directions (changes at $2\Delta^{1/\alpha}$ intervals, for appropriate Δ). The learner has to identify such fast changes, otherwise incurs a pointwise error roughly determined by the margin of η away from $1/2$; this margin is $\mathcal{O}(\Delta)$ (more precisely $C_{\lambda,\alpha} \cdot \Delta$). The slower linear change in one direction ensures that such margin occurs on a sufficiently large mass of points.*

A.3. Proof of Theorem 3

Proof The proof follows information theoretic arguments from Audibert and Tsybakov (2007), adapted to the active learning setting by Castro and Nowak (2008), and to our specific problem by Minsker (2012a). The general idea of the construction is to create a family of functions that are α-Hölder, and cross the level set of interest $1/2$ linearly along one of the dimensions. First, we recall Theorem 3.5 in Tsybakov (2009).

Theorem 8 (Tsybakov) *Let $\mathcal{H}$ be a class of models, $d : \mathcal{H} \times \mathcal{H} \to \mathbb{R}^+$ a pseudo-metric, and $\{P_\sigma, \sigma \in \mathcal{H}\}$ 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.

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 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 $z \in [0,1]$ as

$$f(z) = \frac{z}{2} + \frac{1}{4},$$

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

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

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

For $s \in \{-1,1\}$ and $k \leq K$, and for any $x \in H_k$, we write

$$\Psi_{k,s}(x) = \begin{cases} f(x^{(d)}) + s\frac{C_{\lambda,\alpha}\Delta}{2}, & \text{if } \quad |\tilde{x} - \tilde{x}_k|_2 \leq \frac{\Delta^{1/\alpha}}{2} \\ f(x^{(d)}), & \text{if } \quad |\tilde{x} - \tilde{x}_k|_2 \geq \Delta^{1/\alpha} \\ f(x^{(d)}) + sg(|\tilde{x} - \tilde{x}_k|), & \text{otherwise.} \end{cases}$$

g is such that $g(\frac{1}{2}\Delta^{1/\alpha})) = \frac{C_{\lambda,\alpha}\Delta}{2}$, and $g(\Delta^{1/\alpha}) = 0$. Moreover, it is $\lambda/\alpha^d, \alpha$ Hölder on $[\frac{1}{2}\Delta^{1/\alpha}, \Delta^{1/\alpha}]$ (in the sense of the one dimensional definition of Definition 2) for $C_{\lambda,\alpha}$ small enough (depending only on α, λ), and such that all its derivatives are 0 in $\frac{1}{2}\Delta^{1/\alpha}, \Delta^{1/\alpha}$. Since by definition of $\Psi_{k,s}$ all derivatives in x are maximized in absolute value in the direction $(\tilde{x} - \tilde{x}_k, 1)$, it holds that $\Psi_{k,s}$ is in $\Sigma(\lambda, \alpha)$ restricted to H_k.

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

$$\eta_\sigma(x) = \sum_{k \leq K} \Psi_{k,\sigma_k} \mathbf{1}\{x \in H_k\}.$$

Such η_σ is illustrated in Figure 1. Note that since each $\Psi_{k,s}$ is in $\Sigma(\lambda,\alpha)$ restricted to H_k, and by definition of $\Psi_{k,s}$ at the borders of each H_k, it holds that η_σ is in $\Sigma(\lambda,\alpha)$ on $[0,1]^d$ (and as such it can be extended as a function $\Sigma(\lambda,\alpha)$ on $\mathbb{R}^d$). Finally note that anywhere on $[0,1]^d$, η_σ takes value in $[1/5, 4/5]$ for $\Delta, C_{\lambda,\alpha}$ small enough. So Assumption 2 is satisfied with λ, α, and η_σ is an admissible regression function.

Finally, for any $\sigma \in \{-1,+1\}^K$, we define P_σ as the measure of the data in our setting when $\mathbb{P}_X$ is uniform on $[0,1]^d$ and where the regression function η providing the distribution of the labels is η_σ. We write

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

All elements of $\mathcal{H}$ satisfy Assumption 1.

Let $\sigma \in \{-1,1\}^d$. By definition of P_σ it holds for any $k \leq K$ and any $\epsilon \in [0,1/2]$ that

$$P_\sigma\Big(X \in H_k, \text{ and } |\eta_\sigma(X) - 1/2| \leq \epsilon\Big) \leq (4 - 2C_{\lambda,\alpha})\epsilon 2^{d-1}\Delta^{(d-1)/\alpha}.$$

As $K = 2^{1-d}\Delta^{(1-d)/\alpha}$, it follows by an union over all $k \leq K$ that

$$P_\sigma\Big(X : |\eta_\sigma(X) - 1/2| \leq \epsilon\Big) = \bigcup_{k=1}^{K} P_\sigma\Big(X \in H_k, \text{ and } |\eta_\sigma(x) - 1/2| \leq \epsilon\Big) \leq (4 - 2C_{\lambda,\alpha})\epsilon,$$

and so Assumption 3 is satisfied with $\beta = 1$, $\Delta_0 = 0$ and $c_3 = (4 - 2C_{\lambda,\alpha})$.

Proposition 3 (Gilbert-Varshamov) *For $K \geq 8$ there exists a subset $\{\sigma_0, ..., \sigma_M\} \subset \{-1,1\}^K$ such that $\sigma_0 = \{1, ..., 1\}$, $\rho(\sigma_i, \sigma_j) \geq \frac{K}{8}$ for any $0 \leq i < j \leq M$ and $M \geq 2^{K/8}$, where ρ stands for the Hamming distance between two sets of length K.*

We denote $\mathcal{H}' \doteq \{P_{\sigma_0}, \cdots, P_{\sigma_M}\}$ a subset of $\mathcal{H}$ of cardinality $M \geq 2^{K/8}$ with $K \geq 8$ such that for any $1 \leq k < j \leq M$, we have $\rho(\sigma_k, \sigma_j) \geq K/8$. We know such a subset exists by Proposition 3.

Proposition 4 (Castro and Nowak) *For any $\sigma \in \mathcal{H}$ such that $\sigma \neq \sigma_0$ and Δ small enough such that $\eta_\sigma, \eta_{\sigma_0}$ take values only in $[1/5, 4/5]$, we have:*

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

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) algorithm 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 (2012a) page 10.

By Definition of the η_σ, we know that $\max_{x\in[0,1]^d} |\eta_\sigma(x) - \eta_{\sigma_0}(x)| \leq C_{\lambda,\alpha}\Delta$ (as for any $x, x' \in [0,1]^d$, $\eta_\sigma(x) - x^{(d)}/2 + 1/4 \in [-\frac{C_{\lambda,\alpha}\Delta}{2}; \frac{C_{\lambda,\alpha}\Delta}{2}]$), and so Proposition 4 implies that for any $\sigma \in \mathcal{H}'$:

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

So we have :

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

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

$$\Delta = C_2 n^{-\alpha/(2\alpha+d-1)},$$

as $K = c_3\Delta^{(d-1)/\alpha}$. This implies that for this choice of Δ, Assumption 3 in Theorem 8 is satisfied.

Consider $\sigma, \sigma' \in \mathcal{H}'$ such that $\sigma \neq \sigma'$. Let us write the pseudo-metric:

$$D(P_\sigma, P_{\sigma'}) = \mathbb{P}_X(\text{sign}(\eta_\sigma(x) - 1/2) \neq \text{sign}(\eta_{\sigma'}(x) - 1/2)),$$

where $\text{sign}(x)$ for $x \in \mathbb{R}$ is the sign of x.

Since for any $x \in H_k$, we have that $\eta_\sigma(x) = f(x^{(d)}) + \sigma^{(k)} \frac{C_{\lambda,\alpha}\Delta}{2}$ if $|\tilde{x} - \tilde{x}_k|_2 \leq \Delta^{1/\alpha}/2$, it holds that if $\sigma^{(k)} \neq (\sigma')^{(k)}$ for some $k \leq K$

$$\mathbb{P}_X(X \in H_k \text{ and } \text{sign}(\eta_\sigma(x) - 1/2) \neq \text{sign}(\eta_{\sigma'}(x) - 1/2)) \geq C_4 \Delta^{(d-1)/\alpha} \Delta.$$

By construction of $\mathcal{H}'$ we have $\rho(\sigma, \sigma') \geq K/8$, and it follows that:

$$\begin{aligned} D(P_\sigma, P_{\sigma'}) &\geq \mathbb{P}_X(X \in H_k \text{ and } \text{sign}(\eta_\sigma(x) - 1/2) \neq \text{sign}(\eta_{\sigma'}(x) - 1/2))\rho(\sigma, \sigma') \\ &\geq \frac{K}{8} C_4 \Delta^{(d-1)/\alpha} \Delta \\ &\geq C_5 \Delta \\ &\geq C_6 n^{-\alpha/(2\alpha+d-1)}. \end{aligned}$$

And so all assumptions in Theorem 8 are satisfied and the lower bound follows , as we conclude by using the following proposition from kol (see Lemma 5.2), where we have $\beta = 1$ the Tsybakov noise exponent.

Proposition 5 *For any estimator $\widehat{\eta}$ of η such that $\eta \in \mathcal{P}^*(\alpha, \beta, 0)$ we have:*

$$R(\widehat{\eta}) - R(\eta) \geq C\mathbb{P}_X\big(\text{sign}(\hat{\eta}(x) - 1/2) \neq \text{sign}(\eta(x) - 1/2)\big)^{\frac{1+\beta}{\beta}},$$

for some constant $C > 0$.

In the case $d = 1$, the bound does not depend on α, and the previous information theoretic arguments can easily be adapted by only considering $f(z)$ - the problem reduces to distinguishing between two Bernoulli distributions of parameters $p - \frac{\Delta}{2}$ and $p + \frac{\Delta}{2}$ for $p \in [1/4, 3/4]$.

■

A.4. Proof of Theorem 5

Proof The proof is very similar to the proof of Theorem 3, and thus we only make the construction explicit. Let $\alpha > 0$ and $\beta \in \mathbb{R}^+$.

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

$$K = 4^{-d} \Delta^{(-d)/\alpha},$$

disjoint hypercubes in this grid, and we write them $(H_k)_{k \leq K}$. They form a partition of $[0, 1/2]^d$ that is $[0, 1/2]^d = \bigcup_{k \leq K} H_k$. Let x_k be the barycenter of H_k.

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

$$g(z) = \begin{cases} C_{\lambda,\alpha} 4^{\alpha-1}\Big(\Delta^{1/\alpha} - z\Big)^{\alpha}, & \text{if } \frac{3}{4}\Delta^{1/\alpha} < z \le \Delta^{1/\alpha} \\ C_{\lambda,\alpha}\Big(\frac{\Delta}{2} - 4^{\alpha-1}\big(z - \frac{1}{2}\Delta^{1/\alpha}\big)^{\alpha}\Big), & \text{if } \frac{1}{2}\Delta^{1/\alpha} \le z \le \frac{3}{4}\Delta^{1/\alpha}, \end{cases}$$

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

For $s \in \{-1, 1\}$ and $k \le K$, and for any $x \in H_k$, we write

$$\Psi_{k,s}(x) = \begin{cases} \frac{1}{2} + s\frac{C_{\lambda,\alpha}\Delta}{2}, & \text{if} \quad |x - x_k|_2 \le \frac{\Delta^{1/\alpha}}{2} \\ \frac{1}{2}, & \text{if} \quad |x - x_k|_2 \ge \Delta^{1/\alpha} \\ \frac{1}{2} + sg(|x - x_k|), & \text{otherwise.} \end{cases}$$

Note that g is such that $g(\frac{1}{2}\Delta^{1/\alpha})) = \frac{C_{\lambda,\alpha}\Delta}{2}$, and $g(\Delta^{1/\alpha}) = 0$, and $C_{\lambda,\alpha}$ is chosen such that $\Psi_{k,s}$ is in $\Sigma(\lambda, \alpha)$ restricted to H_k.

Denote $X_1 = (1, ..., 1)$ the d-dimensional vector with all coordinates equal to 1. For $\sigma \in \{-1, 1\}^K$, we define for any $x \in [0, 1]^d$ the function

$$\eta_\sigma(x) = \sum_{k \le K} \Psi_{k,\sigma_k} \mathbf{1}\{x \in H_k\} + \mathbf{1}\{x = X_1\}.$$

Note that since each $\Psi_{k,s}$ is in $\Sigma(\lambda, \alpha)$ restricted to H_k, and by definition of $\Psi_{k,s}$ at the borders of each H_k, it holds that η_σ is in $\Sigma(\lambda, \alpha)$ on $[0, 1/2]^d$ (and as such it can be extended as a function $\Sigma(\lambda, \alpha)$ on $\mathbb{R}^d$ with $\eta(X_1) = 1$). So Assumption 2 is satisfied with λ, α, and η_σ is an admissible regression function.

We now define the marginal distribution $\mathbb{P}_X$ of X. We define p_k for $x \in \mathbb{R}^d$, where we recall that x_k is the barycenter of hypercube H_k:

$$p_k(x) = \begin{cases} \frac{w}{K \mathrm{Vol}\Big(\mathcal{B}(x_k, \frac{\Delta^{1/\alpha}}{2})\Big)} & \text{if } |x - x_k|_2 \le \frac{\Delta^{1/\alpha}}{2} \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathrm{Vol}\big(\mathcal{B}(x_k, \frac{\Delta^{1/\alpha}}{2})\big)$ denotes the volume of the d-ball of radius $\frac{\Delta^{1/\alpha}}{2}$ centered in x_k. This allows us to define the density:

$$p(x) = \sum_{k=1}^{K} p_k(x) + (1 - w)\delta_x(X_1),$$

where $\delta_x(X_1)$ is the Dirac measure in X_1. Note that $\int_{x \in [0,1]^d} dp(x) = \int_{x \in [0,1/2]^d} dp(x) + 1 - w = 1$ as we have by construction $\int_{x \in [0,1/2]^d} dp(x) = w$.

Finally, for any $\sigma \in \{-1, +1\}^K$, we define P_σ as the measure of the data in our setting when the density of $\mathbb{P}_X$ is p as defined previously and where the regression function η providing the distribution of the labels is η_σ. We write

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

All elements of $\mathcal{H}$ satisfy Assumption 1. Note that the marginal of X under P_σ does not depend on σ.

Let $\sigma \in \{-1, 1\}^d$. By definition of P_σ it holds that for any $C_{\lambda,\alpha}\frac{\Delta}{2} \leq \epsilon < 1$:

$$P_\sigma\Big(X : |\eta_\sigma(X) - 1/2| \leq \epsilon\Big) = \bigcup_{k=1}^{K} P_\sigma\Big(X \in H_k, \text{ and } |\eta_\sigma(x) - 1/2| \leq \epsilon\Big) \leq w.$$

and for any $\epsilon < C_{\lambda,\alpha}\frac{\Delta}{2}$:

$$P_\sigma\Big(X : |\eta_\sigma(X) - 1/2| \leq \epsilon\Big) = 0.$$

Thus, in order to satisfy Assumption 3, it suffices to set w appropriately i.e. $w = \mathcal{O}(\Delta^\beta)$. The rest of the proof is similar to that of Theorem 3, where we proceed with $K = \mathcal{O}(\Delta^{-d/\alpha})$, $n\Delta^2 < \mathcal{O}(K)$ which brings $\Delta = \mathcal{O}(n^{-\alpha/(2\alpha+d)})$ and $D(\sigma, \sigma') \geq \mathcal{O}(w) = \mathcal{O}(n^{-\alpha\beta/(2\alpha+d)})$ with σ, σ' belonging to an appropriate subset of $\mathcal{H}$.

■